DOCUMENT RESUME

ED 180 065 EA 012 121

AUTHOR Nasoetion, N.: And Others
TITLE The Development of Educational Evaluation Models in Indonesia.
INSTITUTION United Nations Educational, Scientific, and Cultural Organization, Paris (France). International Inst. for Educational Planning.
PUB DATE 76
NOTE 143p.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Educational Change: Educational Development: Elementary Secondary Education: *Evaluation Methods: Foreign Countries: Item Banks: Learning Modules: Models: National Competency Tests: *Program Evaluation: Teacher Evaluation: Teacher Improvement: Textbook Evaluation
IDENTIFIERS *Indonesia

ABSTRACT

The primary purpose of this project was to develop model evaluation procedures that could be applied to large educational undertakings in Indonesia. Three programs underway in Indonesia were selected for the development of evaluation models: the Textbook-Teacher Upgrading Project, the Development School Project, and the Examinations (Item Bank) Project. The evaluation model project lasted from November 1974 to September 1976. The project directors considered it a success, and the evaluation procedures developed will be used in a series of future development programs. (Author/JM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED180065

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE DEVELOPMENT OF EDUCATIONAL EVALUATION MODELS IN INDONESIA

by

N. Nasoetion, A. Djalil, I. Musa and S. Soelistyo

assisted by

B.H. Choppin and T.N. Postlethwaite

International Institute for Educational Planning (Unesco)

Office of Educational and Cultural Research and Development (BP3K)

of the Ministry of Education and Culture, Indonesia

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

R. Nieto

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA 012 121

Produced and printed in France by the
International Institute for Educational Planning
7-9, rue Eugène-Delacroix, 75016 Paris
September 1976

© Unesco/BP3K 1976

CONTENTS

Preface		5
SECTION I :	INTRODUCTION	7
Chapter 1	Background context - recent educational development in Indonesia	7
SECTION II :	TEXTBOOK AND TEACHER UPGRADING PROJECT	22
Chapter 2	Description of the textbook and teacher upgrading project	22
Chapter 3	Procedures of sampling, instruments, data collection and analysis	29
Chapter 4	Results	35
Chapter 5	General implications	48
SECTION III :	THE DEVELOPMENT SCHOOL PROJECT	51
Chapter 6	Description of the Development School Project in Indonesia	51
Chapter 7	Field trials in 1975 and 1976	58
Chapter 8	Future use of development school evaluation procedures	69
SECTION IV :	THE EXAMINATIONS (ITEM BANK) PROJECT	73
Chapter 9	Models for new examination systems	73
Chapter 10	The try-out	82
Chapter 11	Future use of the item bank in Indonesia	91
SECTION V :	CONCLUSION	96
Chapter 12	General concluding remarks	96

Textbook and Teacher Upgrading Project

Appendix 1	Coding and punching schemes	99
Appendix 2	Summary of item percent changes from pre-test to post-test	101
Appendix 3	Responses on items measuring ability in dictation	104
Appendix 4	Responses on items measuring ability in composition	105
Appendix 5	Responses on items measuring ability in spelling	106
Appendix 6	Mean of selected variables of school facilities by blocks and total	107

Development School Project

Appendix 7	Mean and standard deviations of score and time	108
Appendix 8	Tabulation form	109
Appendix 9	Interview form	110
Appendix 10	General comments on total module	111
Appendix 11	Examples of completed feedback form for module 13	115
Appendix 12	Rasch scaling for item banks	125
Appendix 13	Documentation for an item bank	135
Appendix 14	Specification of computer programmes for item bank operation	137

PREFACE

In 1974 the Government of Indonesia and the United Nations Development Programme (UNDP) approached the International Institute for Educational Planning (IIEP) to provide assistance to the Office of Educational and Cultural Research and Development of the Ministry of Education and Culture (BP3K) in Indonesia in developing evaluation models for certain of the development projects underway or about to start.

Because of its increasing interest in qualitative aspects of educational planning and, hence, in questions of evaluating the development of innovations in education, the IIEP agreed to undertake this task. In November 1974, two staff members of the IIEP, Drs. T.N. Postlethwaite and J.K. Lindsey, visited BP3K for a first round of discussions, as a result of which the following three development projects were selected for the development of evaluation models:

(1) The Textbook - Teacher Upgrading Project.
(2) The Development School Project.
(3) The Examinations (Item-Bank) Project.

The duration of the project was to be from November 1974 to September 1976. Each of the projects were assigned evaluators: Messrs. Aria Djalil and Ibrahim Musa to the first project, Mr. Noehi Nasoetion to the second, and Mr. Sentot Soelistyo to the third. The task was to develop evaluation procedures, test them and revise them so that they would be operative for different grade levels and subject areas in Indonesia.

It was decided to have IIEP staff members visit Indonesia at regular intervals rather than having one person located in Jakarta full time. In this way, it was possible to supply persons with different competencies, as required. Thus, every two to three months an IIEP staff member or consultant visited BP3K for periods ranging from two to four weeks. This proved effective, although occasionally it would have been desirable to have such a person visit BP3K for longer periods of time, in particular for developing operating computer programmes and training educational data processing personnel in Indonesia.

This report provides an account of the development of the various evaluation procedures for each of the three projects. Our expectation at the outset was that this attempt to build evaluation into three educational development projects would result in the utility and necessity of using evaluation feedback data being recognised both for improving development work and for providing educational

decision-makers with a sounder knowledge base for their decisions. As the report indicates, the present evaluation procedures will be used, with some modification, in a series of future development projects. In addition to providing an example of the linkage between evaluation and development in educational programmes, the work on these three projects has also served as an important opportunity for the 'on-the-job' training.

Work on these projects was carried out by the BP3K staff under the direction of Messrs. Aria Djalil, Ibrahim Musa, Noehi Nasoetion and Sentot Soelistyo. Dr. J.K. Lindsey of the IIEP staff worked on the project from November 1974 to September 1975 and Dr. B.H. Choppin (an IIEP consultant) from September 1975 till the completion of the project in September 1976. Dr. T.N. Postlethwaite, an IIEP staff member and from April 1976 an IIEP consultant, also worked on the project from September 1975 to the completion of the project. Funding for the IIEP's participation in the project was provided by UNDP (INS/71/537). All local costs were borne by BP3K.

We should like to thank all of these persons for their dedicated and competent work. In particular, we should like to thank Mr. Moegiadi, Secretary of BP3K, for the exemplary way in which he helped solve so many of the problems which arose in the course of the project, and Mr. Rachmat, the Director of the BP3K research centre and Mr. Soedijarto, the Director of the Curriculum Development Centre, for the unfailing support they gave to all the work.

This project has been one of the first systematic attempts in a developing country to build evaluation into particular educational development projects. This account should be of great interest to educational personnel in many countries of the world.

Setijadi
Director, BP3K
Indonesia

Hans N. Weiler
Director, IIEP
Paris

SECTION I : INTRODUCTION

Noehi Nasoetion

CHAPTER 1

BACKGROUND CONTEXT - RECENT EDUCATIONAL DEVELOPMENT IN INDONESIA

This chapter attempts to describe some of the more recent developments in education in Indonesia. This will provide the reader with a background perspective for the three educational development projects and the development of evaluation models in each of them which are the main concern of this report. It also avoids repetition of facts about Indonesian education in each of the project reports.

CERTAIN CHARACTERISTICS OF INDONESIA

Indonesia is a state of about 13,000 large and small islands. Approximately 6,000 of these islands are inhabited. Indonesia is situated between 6° latitude North and 11° latitude South. The eastern-most part of Indonesia is 5,080 kms. from its westernmost part, that is, from 141° to 95° longitude East. These islands are spread out and near the equator. Indonesia is bordered by the Asian continent to the North and the Australian continent to the South. It is flanked by two great oceans, the Pacific in the East and the Indonesian Ocean in the West. The total land area is approximately 2,000,000 sq. kms., or about 5,000,000 sq. kms. including the sea area. Its East-to-West stretch covers a greater distance than that linking the East and West coasts of the United States of America. Map 1 shows the land area of Indonesia compared to that of Europe. The great number of islands and the distances separating them from each other present communication problems whose solutions have called for, among other things, the launching of the Palapa domestic communication satellite, the upgrading of ports (now numbering 600) as well as airports (there are about 49 capable of accomodating aircraft of various types). The quality of land communication systems - such as highways and railroads - is also being improved.

According to figures released by the Central Bureau of Statistics, Indonesia had a population of 129,082,642 in 1974, including 1,063,960 foreigners. Indonesians grow up under different and complex cultural backgrounds. Each region has its own cultural identity. This phenomenon is brought about by such factors as geographic location, historical background and cultural standards. Each ethnic community has a cultural heritage all of its own. However, these cultural diversities manage to follow a harmonious and orderly course thanks to the application of the 'Bhinneka Tunggal Ika' philosophy, i.e. the philosophy of 'diversity in unity'.

INDONESIA
(compared to Europe)

The majority of Indonesians earn their living in agriculture and plantations. Rice constitutes a major staple food for the people. Tea, coffee, rubber, sugar, oil and logs are some of the country's exports. Indonesia had exported her commodities since the 5th century, especially to India. Her exports to Europe began by the end of the 15th century. Today, relations between Indonesia and the outside world are being continuously strengthened through such forums as the United Nations and the Afro-Asian Conference, the Asean Conference, the Pacific Area Travel Association and the Southeast Asian Ministers of Educational Council. On the educational front, Indonesia is closely co-operating and collaborating with, among others, UNICEF, UNESCO (IIEP), the Ford Foundation, AID, the World Bank, and the Colombo Plan.

EDUCATIONAL DEVELOPMENT 1945-1969

On 17 August 1945, Indonesia proclaimed her independence. From this time on, Indonesians would have to chart their own course and dignity as a nation after being a Dutch colony for 350 years and under Japanese occupation for three and a half years.

Under the political scheme of the colonial rulers education for Indonesians was developed only to a very small extent. Education during the Dutch colonial period sought primarily to provide cheap labour. Those who were fortunate enough to win scholarships received training with the Dutch government agency or with a Dutch businessman, usually as unskilled workers, or second- or third-grade labourers. In 1945 a total of 90 per cent of the Indonesian population remained virtually illiterate. They had no chance to pursue their education because the opportunity was either restricted to or reserved for only certain members of the society.

It is no wonder therefore that soon after obtaining her independence Indonesia opened up unprecedented opportunities for school-age boys and girls to obtain education. As a result, new school buildings were erected and the illiteracy eradication campaign was made a major Government programme. In short, Indonesians were determined to catch up with the progress and development they had missed under colonial rule.

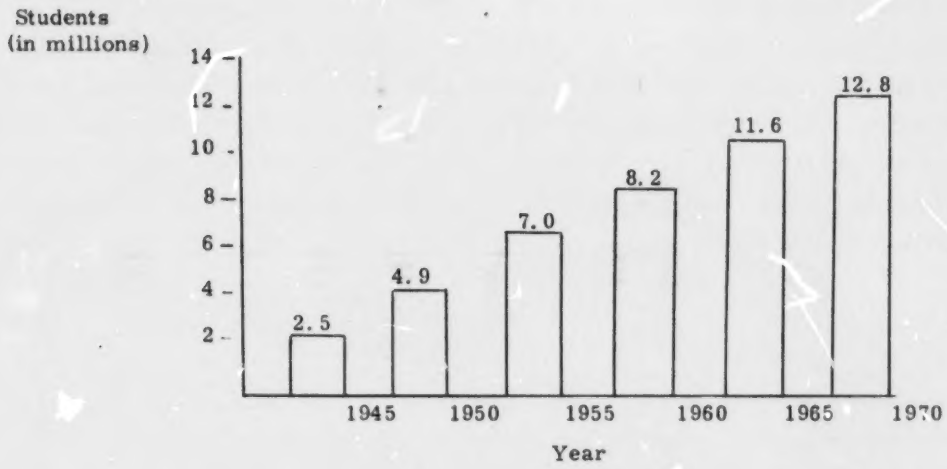
The education sector quickly emerged as an area where fast progress and admirable performance was achieved. One outstanding result of education, of which Indonesia is proud, is the emergence of a national conscience - the awareness of being united as a nation, of a single unifying language and of a culture rich in its variety, diversity and colour. This awareness has since been a unifying force for the nation.

To exemplify the magnitude of the educational development programmes, various aspects of the overall programme are given below, first of all in terms of the development from 1945 until 1968-69 and secondly from 1968-69 onwards. The first Pelita (five-year plan) began in 1968-69.

Expansion of enrolment

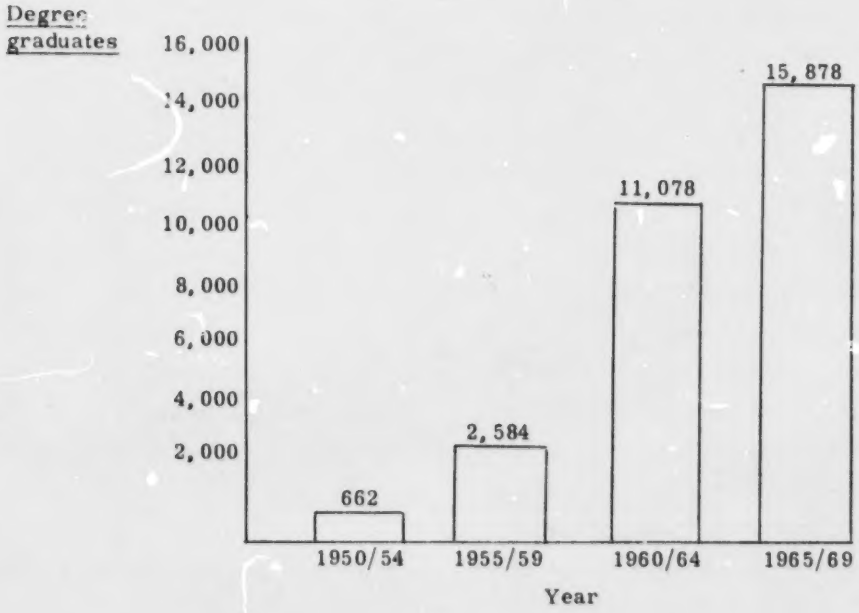
By 1969 primary school enrolment had risen to five times that recorded in 1945, as shown in Figure 1.1.

Figure 1.1 The development of number of students in public and private primary schools (in millions)



Enrolment in junior and senior high schools as well as colleges also increased several times over that recorded in the early years of independence. Figure 1.2 shows the number of full-degree public university graduates excluding IKIP (teachers college) graduates over the period 1950-69.

Figure 1.2 Number of full-degree (M. A.) public university graduates, excluding teachers colleges



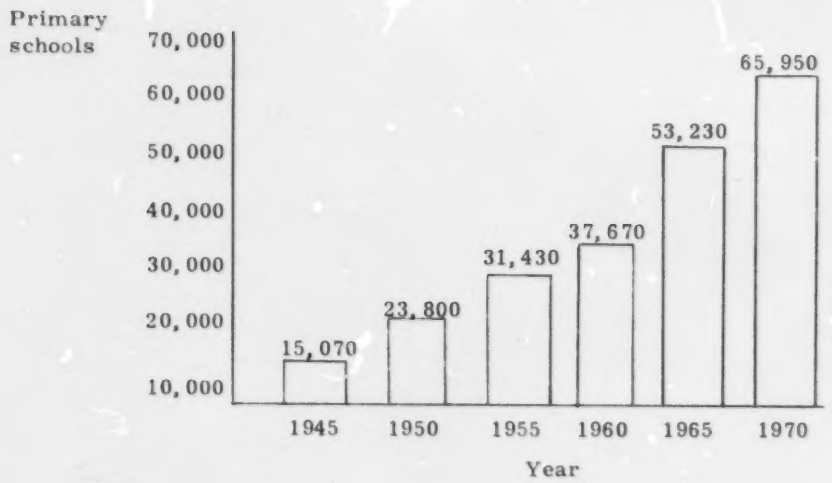
Expansion of school buildings

As the number of school-age children increased from year to year, the Government, through the Ministry of Education, sought to cope with the problem either by increasing the number of students normally allowed for a given classroom or by building new schools. These school buildings are either run by the Government or by private organizations, or are subsidized.

It became clear that in practice educational responsibility did not rest merely with the Government but also with the society as a whole. Community assistance and support are not limited to the moral side. They also extend to the provision of materials and equipment, the erection of new buildings and the supply of school equipment. Many of the newly erected school buildings are therefore the result of joint community efforts.

Figure 1.3 presents the increase in the number of primary school buildings from 1945 to 1970.

Figure 1.3 Number of primary school buildings (1945-70)



Curriculum, textbooks and examinations

The preparation of curriculum and the provision of textbooks have been less speedy compared with the development of school buildings and the rise in school enrolment.

Until 1969 the curriculum remained basically the same as that in use prior to 1945, despite several revisions. The first revision was made in 1957 following the realization that the existing curriculum was no longer up to society's requirements in that it was the heritage of the colonial era. The second improvement took place in 1968 under the name of Curriculum 1968. This improvement was based on educational needs as assessed by educationalists and policy-makers.

Textbooks in use over the 1945-69 period were mostly old foreign books translated into the Indonesian language, with the exception of the textbooks for social studies and the Indonesian language. However, new books have now been written whose contents are in harmony with recent

developments in Indonesia. Not all textbooks were provided by the Government: many students bought their own textbooks. Although the curriculum was applied on a national scale, the textbooks were not. For example, a mathematics textbook used in one particular region or area was not necessarily adopted in another region.

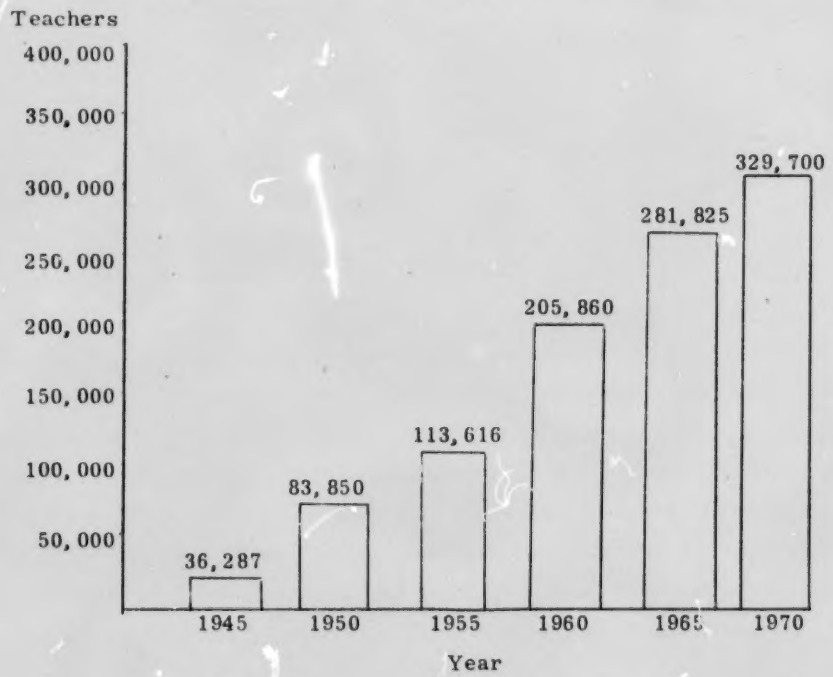
Up to the period preceding the First Five-Year Plan (Pelita I 1969-73), students at primary and junior and senior high schools were required, at the conclusion of their respective courses of study, to take state examinations set on a national level. Examination materials, procedures and norms for determining passes and failures were uniform throughout Indonesia. By 1973 state examinations were gradually abolished and replaced by school examinations. More details on the changes in the examinations system are given in Chapter 9.

Teachers and teaching methods

The large increase in student enrolment over the past 25 years has resulted in an increase in the recruitment of qualified, semi-qualified and provisional teachers. The demand for more teachers has been met by the launching of crash programmes through short-term courses.

There are pre-service and in-service courses offering either written instruction and/or oral instruction. These efforts have been successful in alleviating the problems associated with the shortage of teachers in that an increasing number of teachers is being recruited each year. In order to get a clearer picture of the teacher recruitment development over the 1945-69 period Figure 1.4 presents the increase in the number of both government and private primary school teachers.

Figure 1.4 Increase in the number of primary school teachers in public and private schools



The methods of teaching were more or less uniform and there were also limited facilities available (e. g. teaching equipment, visual aids, laboratories, workshops). The typical teacher employed the lecture method and reduced individual assignments and group discussions to a minimum. The teacher served as the major source of information in the process of learning. Throughout the process of teaching and learning, it was the teacher who played the prominent role. Field work activities were rarely found on the schedule. One result of this has been that the major emphasis of state and school examinations has been on recall - the lowest behavioural category in any taxonomy of educational cognitive objectives.

EDUCATIONAL DEVELOPMENT 1969 TO PRESENT DAY

FIRST AND SECOND FIVE-YEAR EDUCATIONAL PLANS PELITA I AND II (1969-70 - 1978-79)

Problems of education

Over the years preceding the First Five-Year Plan (Pelita I) the Department of Education and Culture set out to identify the problems of education in Indonesia. It was discovered that the problems included qualitative as well as quantitative aspects. These can be categorized into four basic problems:

(1) Obsolescence of the educational system.
(2) Demand for education following the population explosion and the public's growing aspirations and expectations.
(3) Limited sources of labour, facilities and financing.
(4) Necessity for bringing the administration of Indonesian education under one Ministry.

Apart from these four major problems there are a number of other questions whose proper handling is essential to the establishment of more stable conditions for sustained progress in education. They cover, among other things, demography, mass communications media, social and cultural development, economics, politics and security.

Development of activities

The efforts towards overcoming the complex problems of education in general and changes in the educational system in particular involve a good many aspects in other areas and carry far-reaching consequences. The following phasing-up of efforts and/or activities was therefore adopted:

- The period of renovation

At this stage activities concentrated on renewing parts of the teaching force and existing school buildings, and providing improved conditions of work as a starting point for further educational development. Simultaneously, efforts were underway to adapt the system of education to be tailored to public needs and existing national resources. This phase was concluded by the middle of the First Five-Year Plan, i. e. by the end of 1971.

- The stage of consolidation and stabilization

This period saw successful efforts to make the conditions ripe and stable for developments to be able to take place in the real sense of the word. It took approximately two years to bring about such conditions.

- The era of reconstruction

This period covers tangible developments such as an expanding school building programme, new curriculum projects, administrative reorganisation, etc. It went into effect under the Second Five-Year Development Plan (1974-75 - 1978-79). The stages are illustrated in Figure 1.5.

IMPLEMENTATION OF FIVE-YEAR PLANS I AND II

Set out below are the programmes carried out by the Government under the First and Second Five-Year Plans.

Expansion of enrolment

The main objective of this programme is the expansion of enrolment for children of the age bracket 7-12 years, in conjunction with efforts to increase the quality of education. During the first year of Pelita I (First Five-Year Plan), that is from 1969-70, the number of children of the age-group 7-12 years in Indonesia was recorded as 17,838,701. Only 65.6 per cent of this number obtained enrolment in primary schools. The number of children with opportunities for study at school increased a little each year and by the end of Pelita I (1973-74) they represented 66.8 per cent of the total.

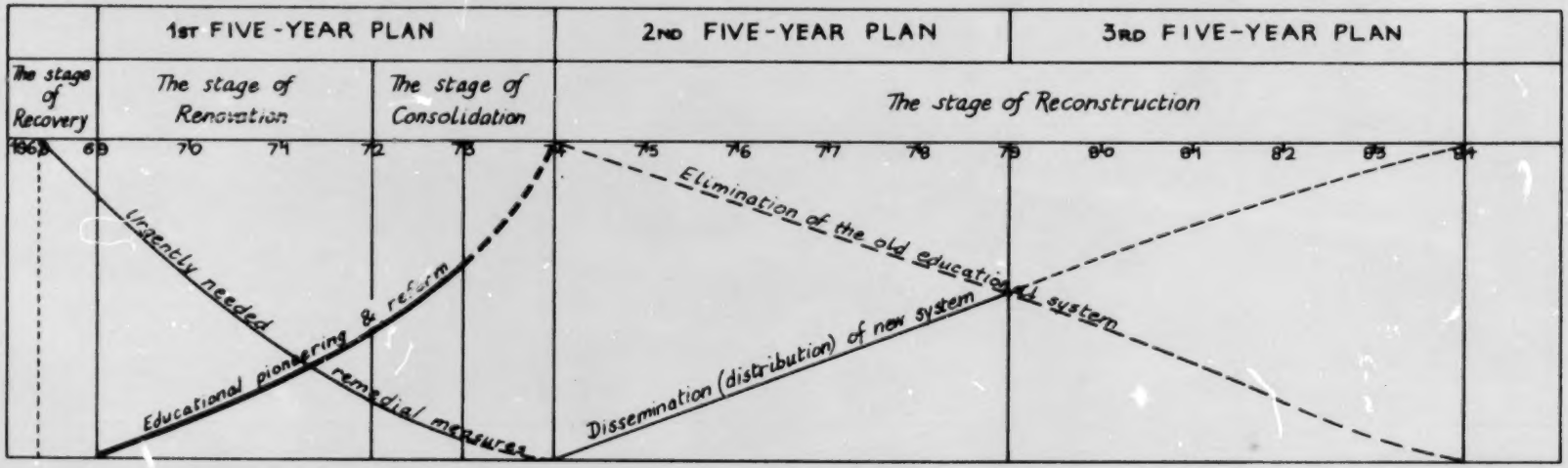
Under Pelita II the Government kept up its efforts to increase the capacity of absorption of the system and the target by the end of Pelita II (1978-79) has been set at 85 per cent (it is estimated that the total number of children in the age-group 7-12 years will reach 23.0 million in 1978.

Steps are also underway to expand gradually enrolment in junior high schools, senior high schools and teacher training colleges, by taking into account the proportional increase in primary school graduates wishing to continue their studies. At the senior high school level the number of students training to be teachers will increase in proportion to the expansion of primary schools requiring more teachers. The attempt to disperse uniformly the learning opportunities available is backed up by the policy of providing scholarships covering all types and levels of education. This scholarship programme is made available in particular to bright students of relatively weak social and economic backgrounds.

Expansion of school buildings

The school building programme also had to expand in proportion to the increased enrolment mentioned above. The existing number of buildings was no longer appropriate given the increase in enrolment. It was therefore decided to continue building new primary schools throughout Indonesia. A programme known as Proyek Instruksi Presiden (Proyek INPRES) was launched to provide funds directly to the Kabupaten (district) for school building programmes.

Figure 1.5 Stages in efforts to solve the problems of education



A total of 6,000 new primary school buildings were built in 1973. The same number was erected in 1974. Then in 1975, a total of 10,000 primary school buildings were commissioned. The same number was constructed in 1976. Their number is to be increased continuously until the target of absorbing 85 per cent of the total number of school-age children has been achieved.

At the secondary school level, all junior and senior high schools, numbering 1, 482 and 518 respectively, have been equipped with science laboratory facilities complete with the necessary equipment. Renovation of all Government-run junior high schools (SMPs) is expected to be completed in 1976. In 1975 a total of 1,000 new classrooms were built with a view to increasing the enrolment capacity of junior high schools. An additional 1,000 new classes and 100 new junior high school buildings were erected in 1976. The construction of new classrooms and new school buildings under the Second Five-Year Plan is to be continued so that a total of 4,000 new classes and 400 new SMP buildings will have been erected by the end of the Plan. The capacity of absorption of teacher training schools and other vocational secondary schools will also be enchanced. Under this programme a total of 13 vocational secondary schools have been constructed and 85 old school buildings renovated in 1974. The same year further recorded the rehabilitation of 16 teacher training school buildings. At the university level this expansion of enrolment capacity had been underway since 1969.

Curriculum, textbooks and examinations

- Curriculum

In 1972 an educational needs assessment survey was conducted in order to identify the aims and objectives for the various levels of education. Based on this survey a master design and a basic course outline for the Development School Project was developed. Module writers developed their modules according to the basic course outline. It is anticipated that this development will have been accomplished by the beginning of the 1980s. The new educational system emanating from the module system is being tried out at eight development schools. However, the Ministry of Education decided as an interim measure to develop a transitional curriculum based mainly on the 1972 national survey and partly on the 1968 curriculum. This was known as the 1975 curriculum.

This new curriculum was introduced on a step by step basis in 1976 in Grades 1 and 4 of primary school, Grade 1 of junior high school and at Grade 1 of senior high school. The programme will call for the application in 1977 of the new curriculum to Grade 5 of primary school, Grade 2 of junior high school and Grade 2 of senior high school. The aim is for all grades to be using the 1975 curriculum by the end of 1978.

- Textbooks

Since the commencement of the first curriculum revision, the need for textbooks, both in terms of quality and quantity, was keenly felt. But then the second curriculum reform managed to overcome the problem of the provision of textbooks and teacher's manuals, at least for primary schools.

For this reason, in the early days of the First Five-Year Plan the Government began initiating a textbook supply programme to complement the efforts aimed at revising the curriculum used by primary schools. As an example, in 1969-70 about 3,230,000 Grade 1 textbooks in mathematics were distributed to primary schools. In the following years millions of science and mathematics textbooks for other grades were distributed to schools throughout Indonesia. These books were based mainly on the 1968 curriculum.

Based on the 1975 curriculum, however, new textbooks for Bahasa Indonesia, social studies, mathematics and science were developed and then tried out in four provinces, i.e. Lampung, West Java, Yogyakarta and South Sulawesi.

Distribution throughout the country took place in 1976 for new textbooks in Bahasa Indonesia, mathematics and social studies for use in Grades 1, 2 and 3. Meanwhile, textbooks for Grade 4 were under revision and textbooks for Grade 5 were being tried out. It is expected that by 1980/81 all primary schools across the country will have received their complete set of textbooks. A similar textbook programme is also underway for junior high schools, senior high schools and vocational schools throughout Indonesia. By the end of 1976 it is expected that 65 per cent of the total of approximately 11,993,500 sets of textbooks for junior and senior high schools will have been printed. Apart from the supply of major textbook items the Government has printed up to 214 titles of library books for primary schools and more than 250 titles for the high schools. These textbooks will operate until the early 1980s when the 1975 curriculum will be merged with the curriculum of the Development School Project.

- Examination system

In 1973 the state examination was abolished and replaced by school examinations. The timetable, procedures and general criteria for the examinations were developed by the Directorate General of Education but blueprints and test items were developed by each school. Thus, different schools measured different learning objectives and each school used its own norm. It was impossible to grade students from various schools in one province onto one scale according to their scores on the school examinations. Therefore students had to take an extra entrance examination if they wished to proceed to the next level of education. It was also impossible for the educational planners to know which school and which area had the weakest achievement either in one province or in the country as a whole. The difficulties experienced in these changes set the scene for the work reported in Chapters 9-11 of this report.

Teachers and teaching methods

Teachers and methods of teaching constitute a factor indispensable to achieving the desired quality of education. In Indonesia there are many 'low-quality' teachers with the result that it is difficult to apply new and effective methods of teaching. This problem has been intensified by the adoption of a curriculum reform and the introduction of other tools of education, such as textbooks and laboratory equipment and other aids, all of which demand more advanced skills and techniques from the teacher as well as up-to-date methods of teaching.

Two alternatives have been chosen to solve the problem of low-quality teachers:

(i) improvement of the teacher training systems through teacher training schools (SPG) and teachers colleges (IKIP);
(ii) provision of in-service training schemes, more popularly known as upgrading ('Penataran'); this programme has been conducted systematically since the early months of the Second Five-Year Plan.

The second alternative is thought to be filling the urgent need more quickly. It involves a step-by-step upgrading of the skills of teachers so that by the end of the Second Five-Year Plan all teachers produced by the old training system will have been upgraded at such workshops. It is worth noting that in 1974 primary school teachers numbered 443,603, junior high school teachers 110,768 and senior high school teachers 60,103, all of which show what an immense effort is required for the teacher upgrading programme.

In order to speed up the completion of the upgrading programme for the teachers a special procedure has been adopted. Primary school teachers are upgraded by Mobile Team Units (MTU) set up in each province. Membership in the Mobile Team Units consists of teachers of teacher training schools (SPG) and area office (Kantor Wilayah) staff trained in Jakarta. Each team deals with a specific subject area. The programme operates Mobile Team Units for science, social studies, mathematics and Bahasa Indonesia. In future it is hoped to have subject (rather than class) teachers in the higher grades of primary school.

A similar scheme is applied for the upgrading of junior high school and senior high school teacher trainers at various national centres (BPGs) such as those in Bandung and Yogyakarta. Those participating in the upgrading programme of the national BPGs comprise IKIP (teacher training college) lecturers, junior and senior high school teachers who have been selected on the basis of a set of special criteria. Those passing the upgrading course with a satisfactory result are then appointed as instructors at regional centres. It is at the national centre (BPG) office that the junior and senior school teachers received their original upgrading instruction. Subjects of the upgrading course generally concentrate on subject matter knowledge, methods of teaching and on how to use teaching aids.

Apart from achieving a standard of better qualified teachers the Second Five-Year Plan also aims at recruiting new teachers for primary school, junior high school and senior high school. It is expected that by 1979 there will be 525,000 primary school teachers (teacher-student ratio 1 : 40, an additional 30,000 junior high school teachers (teacher-student ratio 1 : 18) and approximately 76,000 senior high school teachers (teacher-student ratio 1 : 15).

INSTITUTIONS RESPONSIBLE FOR EDUCATIONAL DEVELOPMENT

The Office of Educational and Cultural Research and Development (BP3K) is a unit under the Department for Education and Culture responsible for the development of education. The decision of the Minister of Education and Culture (No. 079/0 Year 1975) provides that:

- The Office of Educational and Cultural Research and Development (BP3K) has the task of setting up and co-ordinating educational and cultural research activities within the framework of Government policy formu-'ations, of developing different types of educational programmes and long-term planning for and maintenance of the educational system.
- The maintenance of these research and development units in the Department is carried out on the basis of decisions by agencies which are functionally responsible for the implementation of research and development policies.

Figure 1.6 shows the organisational relationship between the BP3K, the Directorate General, the Inspectorate General and the Secretariat General of the Department of Education and Culture.

SELECTION OF THREE PROJECTS

Discussions were held at the Office of Educational and Cultural Research and Development (BP3K) of the Government of the Republic of Indonesia between Drs. Lindsey and Postlethwaite of the International Institute for Educational Planning (IIEP), Paris, and Mr. Setijadi and the senior staff of the BP3K during the period 18-23 November 1974. The aims of these discussions were to select and develop three evaluation projects to be included in the sub-contract to UNDP on educational evaluation. According to the terms of reference of this sub-contract, two of these were to 'test and and apply the methodology of project evaluation' to two development projects, and one to 'assist in the study of students' performance and the examination system' (INS/71/537).

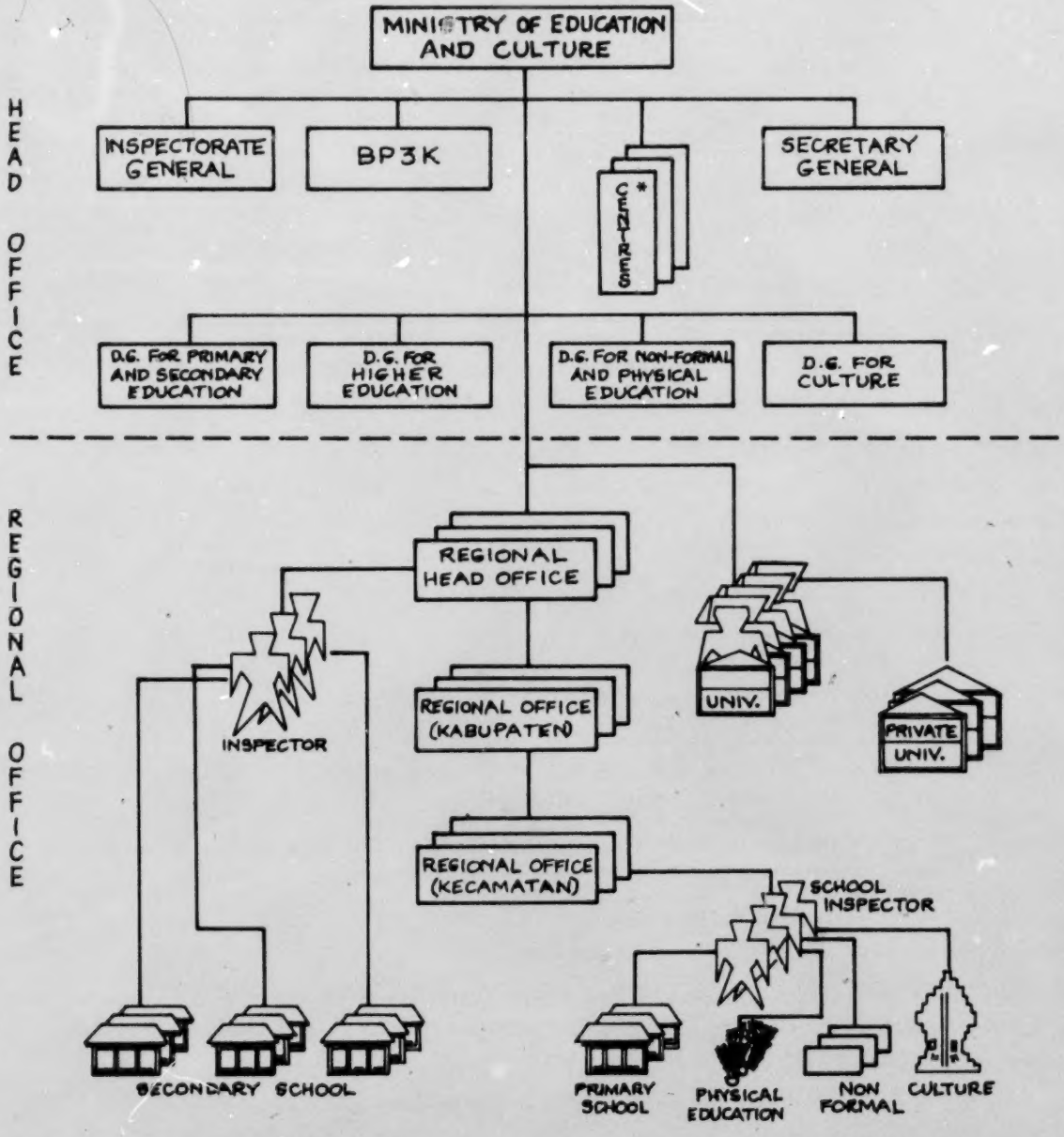
The following large-scale development projects/programmes were reviewed:

(a) Textbook production programme.
(b) Primary school in-service teacher training programme.
(c) Secondary school in-service teacher training programme.
(d) University in-service teacher training programme.
(e) Teacher education (pre-service teacher training) programme.
(f) University consortium project.
(g) University centres of excellence project.
(h) The Grade 6 cross-sectoral national survey of educational achievement.
(i) The radio broadcasting project.
(j) The development schools project, using a new programmed system of modular learning.

Further to this the details of the two types of national examinations, one recently abandoned, the other currently in use, were reviewed, together with suggested possible new systems of national examinations. After this review three projects were selected jointly by the IIEP and BP3K staff. The criteria used were as follows:

(i) The projects should fall within the terms of reference of the UNDP contract.
(ii) They should be able to be completed within some 18-22 months, this being about the time period of the contract.
(iii) They should be high on the priority list of the BP3K both in terms of the utility of the results to the developers of the programmes and the importance of upgrading certain research/evaluation competencies in the BP3K staff.
(iv) They should provide evaluation models which would be of continuing utility to the work of the BP3K.

Figure 1.6 Organizational structure of the Ministry of Education and Culture



*There are six independent centres: Education and training for administrative personnel; Library development; Physical fitness and recreation; Language development; Archeology and material inheritance; and Research on national history and culture. Another four centres are located within BP3K.

The three projects chosen were:

(1) Textbook and Teacher Upgrading Project.
(2) Development School Project.
(3) Examinations Project.

The first project chosen combined projects (a) and (b) in the above list since these two projects are closely inter-related. Textbooks were being introduced progressively into the school system, grade by grade, in the four main subject areas, science, mathematics, Bahasa Indonesia and social studies. In parallel with the preparation and introduction of the textbooks, teachers were being upgraded to be able to use them. Unfortunately, this project had been in operation for some time, so that it was necessary to start the evaluation at a relatively high grade level (Grade 5). However, since such a large portion of the education budget was being allocated to this project, it was felt necessary to introduce evaluation of it as soon as possible.

Because of the high priority given to the Development School Project by the BP3K staff, this project was also chosen for evaluation. The testing of the modular learning system in eight experimental schools was to start at the beginning of 1975, and a proper evaluation was considered to be essential.

The third project, it was felt, should undertake some basic work which would help in launching a more satisfactory national examination system for future years.

SECTION II : TEXTBOOK AND TEACHER UPGRADING PROJECT

Aria Djalil and Ibrahim Musa

CHAPTER 2

DESCRIPTION OF THE TEXTBOOK AND TEACHER UPGRADING PROGRAMME

BACKGROUND

The various curriculum reforms and the various agencies responsible for them have been described in Chapter 1.

The Textbook and Teacher Upgrading Project is a large one in that the final aim is to supply textbooks free of charge to all children at each grade level of primary school in four subject areas, mathematics, science, social studies and Bahasa Indonesia, and to upgrade all primary school teachers. In 1973 there were 13.6 million primary school students (57 per cent of all children eligible for primary education) and 425,000 teachers. By 1980 it is hoped that 85 per cent of all children eligible for primary education will be enrolled in school. The book production and teacher upgrading budget is estimated to be approximately US$40 million, which is allocated for in-service training for primary school teachers, regional supervisors and tutors, (a total of about 400,000 people) and for producing and distributing 179 million books.

There are two major types of textbook in Indonesia. The first textbook (referred to as the 'old textbook' in this report) was produced by many different authors over the period 1945-68. A second textbook (the 'new textbook') was produced by the Curriculum Development and Teaching Methods Project (known in Indonesia as PKMM) during the period 1969-73 and later by the Project for the Improvement of Primary Schools (P3D) during the period 1973-80. P3D is also responsible for the teacher upgrading, i.e. short courses to train teachers to teach the new textbook and to use the teachers' guides.

Differences between old and new textbooks

The old textbooks were produced by many different authors with different tastes and styles, and were of very varying quality. Furthermore there was no one curriculum 'objectives blueprint' worked out for any one subject matter at any one grade level. The old textbook also tended to be subject-matter-oriented, e.g. social studies was not an integrated subject but tended to be taught separately as civics, history and geography. The same was true for science (i.e. biology, chemistry, physics).

When P3D assumed the responsibility for this task, standard blueprints were constructed and every effort was made to have appropriate horizontal and vertical integration of topics and

subjects. In certain cases extra materials were added, and certain methods of teaching were revised. In Bahasa Indonesia, for example, the new textbooks included the structural analysis and synthetic approach (SASA).

Thus, the aim of producing this new type of textbook was an attempt to establish materials of equal quality throughout the country - a very different situation from the previous era with a differing quality of books in different regions and schools.

Try-out of new textbooks

The Project Implementation Unit (PIU) of P3D is responsible for trying out the new textbooks - including the teacher upgrading in try-out schools.

The initial step is to form subject matter teams which are responsible for writing the textbooks and deciding on certain teaching methods, and for the evaluation procedures for the try-out. The subject matter teams are also responsible for planning and executing the upgrading of teachers to be used in the try-out.

The required number of books for the try-out are printed, but as a result of the try-out (pre- and post-test scores, and teachers' comments), the manuscripts are then revised by the textbook writers and subject matter specialists (e. g. university teachers).

The upgrading programme

The upgrading programme is not only for subject matter (no longer class) teachers, but also for school principals and subject matter supervisors. This is a three-week programme, the first two weeks of which (attended by teachers and supervisors only) consist of:

(i) the textbook material and teaching and learning methods in each subject matter at each grade level;
(ii) the production and use of teaching aids;
(iii) general principles and procedures of the evaluation of student learning;
(iv) some simulated teaching practice.

In the third week the school principals attend the course and this third week is concerned mostly with techniques of dealing with children's questions and problems, observation of children's progress, test administration, etc.

This upgrading for try-out takes place just before the beginning of each academic year. Table 2. 1 presents the upgrading for try-out in the period 1973-76.

Table 2. 1 Upgrading of try-out teachers, 1973-76

1973-74	1974-75	1975-76
1. Grade 3 Bahasa and social studies	Grade 4 Bahasa, maths, social studies and science	Grade 5 Bahasa, maths, social studies and science
2. Grades 1, 2, 3 for maths.		

P3D is responsible for the evaluation of its upgrading courses, but it has proven impossible to solicit complete information from P3D about either the nature or the results of the evaluation.

P3D is also responsible for the evaluation of the textbooks in their development stage. A pre- and post-test is given to the students in the try-out schools at the beginning and end of each quarter. Teachers give their comments on the new textbooks through questionnaires and once a month each class trying out the new textbooks is observed by the subject matter supervisors.

On the basis of the test results and other information, the new textbooks are revised. Expert opinion is also used. Priority has been given to the feedback for the revision of the new textbooks and the feedback for improving teacher (and supervisor and school principal) upgrading has received little effort. This was the P3D formative evaluation.

Implementation

When it comes to the implementation of the new textbook programme into all primary schools in Indonesia, mobile teams are used for the massive teacher upgrading programme.

Selected persons who will form the mobile teams are trained in Jakarta by those who upgraded the try-out teachers. After training they are formed into mobile teams to serve the various provinces in Indonesia - one mobile team per province. The mobile team members are selected mostly from among the teachers of training colleges for primary school teachers (SPGs) and to some extent the teaching staff of the institutes of teacher training and educational sciences (IKIPs) and from among senior primary school teachers. This second wave of upgrading teams is located in the various provinces of Indonesia. The content of upgrading for the mobile teams is very similar to that for the try-out teachers. Apart from the subject matter, content emphasis is given to the writing of training packages which can then be used in the courses in the various provinces.

The timetable for trying out books and introducing them to all schools is presented in Table 2.2.

Table 2.2 Timetable for trying out books and their introduction

Time	Try-out	Introduction in schools
1 January 1974	Grade 3 Bahasa, Grades 1, 2 and 3 maths, Grade 3 social studies	-
1 January 1975	Grade 4 Bahasa, maths, social studies and science	Grade 1 Bahasa 1/
1 January 1976	Grade 5 Bahasa, maths, social studies and science	Grade 1 maths, Grade 2 Bahasa 1/, Grade 3 social studies
1 January 1977	Grade 6 Bahasa, maths, social studies and science	Grade 2 maths, Grade 3 Bahasa, Grade 4 science
etc. until 1980-81	-	etc. until 1980-81 when all books will be available at schools

1/ Grades 1 and 2 Bahasa were tried out before P3D was established.

THE BP3K-IIEP PROJECT

BP3K was asked to evaluate the programme in terms of whether more teacher upgrading improved student learning and whether the new textbooks were 'better' than the old. It was at this point that the IIEP was asked to help in the evaluation work.

The task was to help in producing and trying out a system of evaluation for both the in-service training and the textbooks. In November 1974 it must be recalled that it was difficult for the programme developers to know if any observable effects on student performance were due to the new textbooks or to the teacher upgrading: little evaluation of the teacher training course had been made in terms of teaching methods learned either at the end of the course or how they were used later in the classroom. Nor had any detailed item analysis using formative tests been made to determine if the textbooks were biased in terms of sex, social group, urban/rural, etc. differences. Hence, this evaluation was different from, and in a sense complementary to, the formative type of evaluation typically undertaken by P3D.

DESIGN

Evaluation design - First plan

Bahasa Indonesia was to be tried out in 65 schools in five regions.

As a pilot project, Grade 5 Bahasa Indonesia was chosen since, as can be seen from Table 2.2, it was to be tried out in 1976 and implemented in 1978. Hence there was the possibility of revising the textbook on the basis of both the P3D try-out and the BP3K try-out.

The first part of the project was to involve an evaluation of the teacher training at the end of the two-week course to determine if the teachers had actually mastered the three general objectives desired: ability to specify curriculum objectives from a textbook, to construct and administer formative tests, to analyse the data and write the reports. They must also master certain teaching methods, create their own instructional aids and practise certain teaching techniques.

The second part of the project would involve evaluation of the teacher training and the textbook in use in the pilot schools two months after the two-week upgrading course. Some 24 of the 65 schools would be chosen, and teachers from an additional 24 schools also upgraded in the same teaching methods. The research design is as follows:

A : Teacher training + New Textbook	C : New Textbook only
B : Teacher Training only	D : No Teacher Training and Old Textbook

with 24 teachers (and their classes) involved in each block. Thus Block A, the 24 schools chosen from the 65, proceeds as would normally be expected. The 24 additional teachers upgraded, in Block B, would be sent back to their schools to teach not using the old textbook but the new methods.

Two other sets of 24 teachers would also be chosen, one as a control, for observation only (D) and one to whom the new textbooks would be given without the upgrading, so that they will use the traditional teaching methods. The classes in each block would be matched as closely as possible. An observation schedule would be prepared for detailed observation of the first block and for periodic observation of the other three blocks. A summative test would be prepared to test the objectives of the curriculum and would be administered to all of the students involved along with a home background questionnaire. The data would be analysed and reports written to show possible revisions of the textbooks, as well as possible improvements in the teacher upgrading in preparation for national implementation the following year.

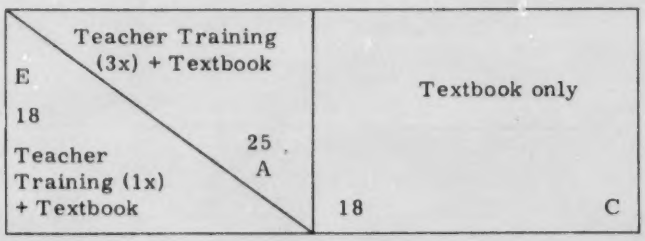
Evaluation design - Second plan

As time moved on complications began to emerge. Some teachers were being upgraded by P3D (in Bahasa Indonesia at least) three times, once for Grade 3, once for Grade 4 and once for Grade 5. It will be recalled that for the purpose of this study other teachers were only upgraded once by the P3D team, whilst others were not upgraded. Later on it was discovered (in the third plan, see below) that local upgrading was taking place (i. e. by others than the P3D upgraders).

It is perhaps worth interjecting at this point that classical research designs such as the one given in the first plan stand a very low probability of working in the field of education - simply because of the rapid changes in education such that a block in such a design as above may well change its characteristics over time (as happened here) and hence blocks may become confounded.

To return to the second plan, because of constraints of lack of money and personnel it was thought that it might be necessary to restrict the number of regions in which the evaluation exercise might be conducted. However, in May 1975 it was suggested that the original design be modified to take care of the thrice-upgraded teachers in the following way:

Create a new block E which would also have both teacher training and the new textbook, but which would consist of schools not originally in the try-out groups. These would then have only one upgrading, etc. This would be done by reallocating schools already chosen for blocks B, C and D.



Final evaluation design

Two regions were chosen in which to conduct the work: Lampung (in Southern Sumatra) and Kuningan (in West Java). These are not too far from Jakarta, in terms of conducting the data collection. Furthermore, since the criterion test was one of Bahasa Indonesia, it was necessary to examine the

effects of speaking the local language only, that and Bahasa Indonesia, or Bahasa Indonesia only. Lampung children tend to speak a local language and Bahasa Indonesia, and Kuningan children tend to speak the local language only.

It was decided to examine and compare schools using the old textbook and schools using the new one. Since it also occurred that some teachers had no upgrading, some one upgrading course, and a third group three upgrading courses (specialist teachers of Bahasa Indonesia having had an upgrading course for Grade 3, one for Grade 4 and one for Grade 5).

Figure 2.1 shows the way in which these two sets of variables (textbooks and teachers) are distributed for the experiment.

Figure 2.3 Types of teachers and textbooks in research project

	Upgrading		
	None	Once	Three times
Old textbook	D 8 schools	B 8 schools	
New textbook	8 schools E	8 schools C	8 schools A

In each cell there are eight schools - four have been selected from the Lampung area and four from the Kuningan area. Only in those schools with the new textbooks has there been the possibility for some teachers to be upgraded three times.

- Textbooks and teacher upgrading

Students were to be tested in Bahasa Indonesia with a pre-test and post-test over the first trimester (January to April 1976). The Raven's Progressive Matrices test was also to be given at the beginning of the trimester. The evaluation team selected this to be used as a possible covariate in the analyses. In any case, it was proposed to undertake an analysis of covariance to examine the differences between the cells. The test in Bahasa Indonesia was divided into two parts: one part consisting of items measuring objectives common to both the old and new textbooks, and a second part of items measuring objectives only in the new textbook. Comparisons were to be made between all cells separately on these two scores (common and new textbook only).

When there is wide variation between schools within cells on the post-test total scores, information on differences between schools, such as student background, teacher background and teaching procedures (from observers) will be given to the organisers of the project.

- Feedback on new textbook

It was agreed to undertake an item analysis on the post-test results taken by the students in the 24 schools having used the new textbooks. Where items were poorly answered (say less than 70 per cent of the students answering the item correctly) or there was a wide range of scores between

schools (say 40 per cent or more), these were to be identified and the results fed back to the textbook writers, together with information collected bearing on the possible reasons for such low or wide achievement.

- Feedback on teacher upgrading

In 1975 an upgrading course was held for twelve teachers from Kuningan and twelve teachers from Lampung. The course was approximately 80 instructional hours in length. A test was devised to be used as a pre- and post-test. Unfortunately, the items did not cover all the general objectives of the upgrading course, namely:

(a) Principles and methods of the structure-analysis-synthesis approach (SASA) for the teaching of Bahasa Indonesia.
(b) Mastery in the use of teachers' guides.
(c) Formulation of instructional objectives.
(d) Aesthetic appreciation.
(e) New spelling.
(f) Use of the language guide for the practice of teaching.

Furthermore, the quality of some of the items (for validity, unambiguity, objectivity, etc.) was dubious. However, on this 39-item test the average pre- and post-test scores were:

	Pre-test	Post-test
Kuningan	17.1	22.5
Lampung	11.3	22.0

Again, no item analysis was undertaken so that no detailed feedback was given on which objectives were well or poorly attained.

However, the main difficulty was that the persons responsible for the in-service training were unable, at least up to December 1975, either to give the evaluators a curriculum blueprint for the in-service course or to give a simple list of operationalised (or even unoperationalised) objectives. Hence the 1975 work on teacher upgrading had been of no use. In an attempt to do some sort of evaluation for teacher upgrading, a limited and simple plan was devised for 1976. By this time it was too late to evaluate the upgrading per se since it was finished and the teachers were already in the schools.

It was therefore decided to reappraise the objectives of the teacher upgrading course and select those stated by the teacher upgraders to be most important. These objectives (formulated as teacher behaviours) were to be introduced into an observation schedule instrument and the observers would observe the extent to which these behaviours were being used by all teachers. Feedback would be given to the upgrading course organisers on:

(a) the difference in teaching behaviours between teachers with no upgrading, one upgrading and three upgradings;
(b) the difference in students' test scores according to whether they were under teachers with no upgrading, one upgrading or three upgradings.

This is clearly not as much as one would wish, but given the collapse of the 1975 work, this was as much as one could hope for in 1976.

CHAPTER 3

PROCEDURES OF SAMPLING, INSTRUMENTS, DATA COLLECTION AND ANALYSIS

SAMPLING

P3D, when trying out its textbooks and teacher upgrading programmes, had selected 4 provinces out of the 26 in Indonesia. Two of the provinces were in Java (West Java and Yogyakarta), one in Sumatra (Lampung) and one in Sulawesi (South Sulawesi). It should be noted that 80 per cent of the Indonesian population live in Java and Sumatra. Within each province four wilayah (or sub-districts) were chosen at random, and within each wilayah three schools were selected, one of high, one of medium and one of low standard. The P3D criteria for standards were:

(a) the building and equipment of the school;
(b) the teachers' qualifications;
(c) the teacher-student ratio;
(d) the teacher-class ratio.

Given the financial, temporal and manpower constraints of the BP3K, only two districts were chosen for the BP3K evaluation study. These were South Lampung (Lampung) and Kuningan (West Java). Wilayah were then drawn at random. Within wilayah five schools were drawn to fit with the blocks A, B, C, D and E. Schools were checked to see whether:

(i) they still existed and if it was reasonable to visit them in terms of communication problems and the amount of time required to get to them;
(ii) the location of the schools in the blocks were not too close to one another, thus reducing contamination to a minimum;
(iii) they had a minimum of ten students in Grade 5.

There were some difficulties in finding replacement schools when the above criteria were not fulfilled and hence, as can be seen from Tables 3.1 and 3.2, not all wilayah were represented by one school for each block in the final sample. These schools are a random sample of the schools in Lampung and Kuningan. Thus, a random sample was selected of 40 schools from 8 wilayah in two provinces (primary schools in Indonesia total 65,950, scattered over 2,356 wilayah in 26 provinces).

Nevertheless, the sample in Lampung might be considered to represent provinces where people speak a mixture of the vernacular and Bahasa Indonesia. The regional language that is widely used is Javanese, but the medium of instruction in schools is Indonesian. Since the population in Lampung is a mixture of several tribes - among others are Javanese, Sundanese, Lampungnese - Indonesian is often used as the medium of communication among them. The schools

Table 3.1 Schools in the BP3K/IIEP study in the province of Lampung

No.	Sub-district	Name and status of PSs	Status/block A	B	C	D	E
I	Gedung Tataan	1. Karang Anyar	X				
		2. Bagelen					X
		3. Kurungan Nyawa		X			
		4. Bogorejo			X		
		5. Kutoarjo			X		
		6. Negeri Sakti				X	
II	Pringsewu	7. Pringsewu I	X				
		8. Pajaresuk		X			
		9. Ambarawa I		X			
		10. Podomoro					X
		11. Sidoharjo				X	
III	Kota Agung	12. Kota Agung III				X	
		13. Kota Agung IV				X	
		14. Way Maja					X
		15. Kalisari	X				
IV	Penengahan	16. Tetaan	X				
		17. Pasuruhan II			X		
		18. Penengahan			X		
		19. Rawi		X			
		20. Pasuruan I					X
	TOTAL (4)	(20)	4	4	4	4	4

Table 3.2 Schools in the BP3K/IIEP study in the province of West Java (Kuningan)

No.	Sub-district	Name and status of PSs	Status/block A	B	C	D	E
I	Kodu Gede	1. Bayuning I			X		
		2. Haur Kuring II		X			
		3. Jembar II				X	
II	Cidahu	4. Cieruih	X				
		5. Cikidung Girang	X				
		6. Cimulya			X		
		7. Kertawinangun		X			
		8. Partawangun					X
		9. Cidahu				X	
III	Ciwaru	10. Karang Kencena			X		
		11. Cilayang		X			
		12. Sukajaya					X
		13. Ciwaru				X	
IV	Jalaksana	14. Sadamantra	X				
		15. Citapen	X				
		16. Singkup			X		
		17. Pesing		X			
		18. Pajadanu					X
		19. Manislor					X
		20. Gara Tengah				X	

selected are usually located more than 40 kms. away from the capital of the province in agricultural areas. The majority of students come from relatively low income homes (but a little bit higher than Kuningan).

Kuningan, on the other hand, represents a province with almost all people speaking vernacular Sundanese. The medium of instruction in schools in Grades 1 to 3 is, for the greater part, Sundanese. The schools are located about 200 kms. from the capital of the province, and as in Lampung, schools in Kuningan are in agricultural areas, and students generally come from low income homes.

INSTRUMENTS

Instruments for the evaluation of the textbooks

A battery of instruments was produced for this purpose:

(a) Student test of Bahasa Indonesia (used both as pre- and post-test) consisting of 60 items, to be completed in 90 minutes. Items could be divided into two sets, i. e. 49 items common to the old and new textbooks, and 11 items for the new textbook only. The 60 items measure different aspects of Bahasa Indonesia as follows:

Aspects	No. of items
1. Different aspects of grammatical structure	19
2. Reading comprehension	5
3. Dictation	5
4. Composition	3
5. Vocabulary	23
6. Spelling	5

This test reflects the curriculum of the first trimester of Bahasa Indonesia in Grade 5. All items were pre-tested in the fall of 1975 and, where necessary, were revised. Given the small number of items for anyone sub-score, it was decided that sub-score totals would not be calculated. Only two scores were calculated: one for the 49 items common to the old and new textbooks, and one for the 11 items for the new textbook only.

This test was constructed by a team of test writers (from IKIP Bandung) from a blueprint based on a content analysis of the old and new textbooks. The sources for this test were books used by non try-out schools and the new book itself. The method of measuring the test validity was through a pre-test, as mentioned above, and some revisions were necessary.

Each test item was scrutinized for content relevance and judgements were made on whether or not it was a well written objective item.

(b) Teacher comment form: The 24 teachers using the new textbook (Blocks A, C and E) were asked to respond to a series of general questions about the new textbook and to indicate specific difficulties encountered with specific parts of the book. This information was expected to be particularly helpful for supplying qualitative information about poorly achieved objectives.

Instrument for evaluating the upgrading programme

An observation schedule (22 items) was developed. Each classroom was observed once a month by a trained observer. The observers noted and judged such matters as teacher formulation and clarity of instructional objectives, lesson plans, requested feedback, use of feedback, how a new topic was introduced, teaching aids used, exercises given, homework given, teaching activity (passivity, discipline, teacher questioning and explaining behaviours, etc.). A rough estimate was also made of the amount of time used for such things as class management, announcements, teaching the subject, individual or group exercises, questions and answers, evaluation, etc. These teaching behaviours were selected by the evaluation team based on their knowledge of the upgrading programme since it was impossible (as mentioned before) to have the upgraders state the specific objectives of the upgrading course.

These data will be used in two ways:

(a) to compare upgraded (once and thrice), and non-upgraded teachers in the exhibition of these behaviours;
(b) as qualitative data characterizing schools to be examined for their association with achievement differences between schools in Bahasa Indonesia.

Other instruments

- The Raven's Progressive Matrices test (RPM) 60 items (45 minutes)

This is basically a test of non-verbal ability, measuring general intelligence of the student of age 10 and above. The RPM score was used as a covariate in the analysis.

- Teacher questionnaire (currently 13 items)

This questionnaire collected information on such variables as sex, age, teacher qualification, years of teaching experience, teaching subject preference, type of upgrading course attended, etc. These data were not used in the quantitative analysis, but to serve as a general description of teachers' backgrounds for a given school or a group of schools. They were also used for characterizing schools where there were large differences in score.

- School facilities questionnaire (37 items)

This questionnaire was designed to collect information on the state of the school building, classrooms, staff room, playground, the type of community the school serves, how many children at one desk, the health conditions in school, etc. These data were not used in the main analysis but it was decided to produce univariate statistics (mean, standard deviation and frequency distribution) for each variable. It was not intended to plot this information systematically against achievement scores but when inspecting large differences between blocks (or schools), then it can be seen to what extent any aspect of school facilities is related to any differences in Bahasa Indonesia achievement.

- Student questionnaire (20 items)

This questionnaire was designed to obtain information on sex, age, amount of grade repeating, major language of home and community, illness of children, amount of homework and where and when undertaken, type of lighting in home, father's occupational status, father's ability to pay school fees, who takes care of the child, number of persons in household, student reading habits, etc.

Some of these data were used for breakdowns in the item analyses, but some were used as covariates in the analysis of covariance.

DATA COLLECTION

Training of data collection personnel

To collect the data in the schools, the evaluation team was assisted by four senior people of the Regional Office of the Ministry of Education and Culture at the provincial level (two from Lampung and two from Kuningan).

They were given training to:

(i) administer the RPM and the pre- and post-tests of Bahasa Indonesia;
(ii) carry out observation and complete the school facilities questionnaire;
(iii) administer student and teacher questionnaires and instruct teachers on the completion of teachers' comments forms.

This training was also given to 20 school supervisors at sub-district level (9 people in Lampung and 11 in Kuningan) to help carry out the pre- and post-testing.

Forty teachers (20 in Lampung and 20 in Kuningan) were given training on how to fill in the teachers' and students' background questionnaires. This training was given in January 1976 for two days.

Data collection

The first stage of data collection involved the pre-test and RPM, students' and teachers' background questionnaires, school facilities questionnaire, teachers' comments on the new book and observation of teaching behaviour. Except for the pre-test and RPM data collected in January 1976, other data were collected in stages, so that everything was finished towards the end of the first semester.

Some difficulties were encountered during the collection of data on the students' backgrounds, especially in filling out the parents' occupation and information on learning habits. The students were asked to write down their parents' occupation clearly so that the evaluation team could classify them into low, medium and high categories of status (in terms of social and economic status). Not many Grade 5 students were able to write down their parents' occupation, although the teacher gave directives. To overcome this problem, the evaluation team and supervisors examined each questionnaire the students handed in. When items were incompletely or incorrectly filled in the questionnaires were returned to the school for further details. The answers to the question on 'learning habits' indicated that the students had not fully understood the question. It was impossible to correct this error. Another difficulty arose in the observation of teacher behaviour. The observations that

should have been done four times in each school (once a month, in one semester) could only be done three times as the supervisors assigned could not be released fully from their duties at the Regional Office.

The second stage of data collected were post-test data. This exercise went smoothly. Only 15 students in Lampung and 19 in Kuningan who took the RPM and pre-test were not able to take the post-test. This was because at the time the post-test was given the children were helping their parents in the rice fields (harvest time), or because they were ill, or because they had dropped out. The total number of students to take the RPM, pre-test and post-test was 542 in Lampung and 465 in Kuningan.

Return of data

The pre-test, RPM and post-test data could be taken to Jakarta directly because at each testing session there was a member of the evaluation team who went to Lampung and Kuningan. Almost all data on the teachers' background and school facilities were sent to BP3K at the end of January 1976. Teachers' comments and students' background data were sent to Jakarta every month by the supervisors.

Data coding and punching

The coding was done by an evaluation team with five assistants who transferred the data onto code sheets. Punching was done by the Soedarpo Corporation in Jakarta, whereas the data processing was done by the BP3K Wang computer evaluation team.

The coding and punching schemes are given in Appendix 1.

DATA ANALYSIS

As mentioned earlier, two types of data were used in the analysis:

(a) The first type used in the quantitative analysis consists of the pre-test and RPM, post-test, type of textbook, number of upgradings and some of the student background variables.
(b) The second type used in the qualitative analysis consists of the background of teachers, school facilities, observation results and teachers' comments on the new textbooks.

The quantitative data were processed using the Wang computer and consisted of:

(a) An item analysis (percent replying to each response and point bi-serial correlations between each item - and each response - with total score minus that item).
(b) Product moment correlation matrices of the criterion and 19 other variables.
(c) Analysis of variance.
(d) Analysis of covariance.

The qualitative data were processed manually to obtain:

(a) Mean and standard deviation for each variable in the teacher background and school facilities questionnaires. These were calculated by block and also by regions within blocks.
(b) Frequency distribution of teacher behaviour variables, by regions within blocks and by block.
(c) Summary of the teachers' comments on the textbooks by block.

CHAPTER 4
RESULTS

This chapter presents the major results of the study as well as selected minor results which are thought to be of importance.

GAINS IN ACHIEVEMENT IN SEMESTER 1

Table 4.1 presents the gains in Semester 1 in Bahasa Indonesia from the pre-test to the post-test.

Table 4.1 Average score changes from pre-test to post-test by region and block

	Old Textbook				New Textbook						Total	
	B		D		A		C		E			
	X	SD	X	SD	X	SD	X	SD	X	SD	X	SD
Region 1 (Lampung)												
Pre-test	21.65	5.12	19.88	5.74	23.39	5.84	18.00	5.13	17.88	5.03	20.02	5.38
Post-test	24.40	6.77	23.46	7.49	31.42	6.76	23.89	7.04	22.75	6.84	25.10	7.01
Gain	2.75	-	3.58	-	8.03	-	5.89	-	4.87	-	5.08	-
Region 2 (Kuningan)												
Pre-test	14.00	4.28	18.61	4.94	23.16	5.68	20.49	6.38	16.63	3.82	18.47	5.09
Post-test	13.33	5.75	26.48	6.82	25.94	6.57	26.48	7.02	29.32	5.06	28.03	6.28
Gain	17.33	-	2.87	-	2.78	-	5.99	-	12.69	-	9.56	-
TOTAL												
Pre-test	17.90	4.75	19.31	5.42	23.27	5.79	19.03	5.71	17.32	4.55	19.31	5.25
Post-test	27.71	10.71	24.83	7.76	28.63	8.03	24.98	7.76	25.77	9.71	26.46	8.98
Gain	9.81	-	5.52	-	5.36	-	5.95	-	8.45	-	7.15	-

It would be useful to recapitulate the definition of blocks at this point:

Block A = New textbook - teachers with 3 upgradings - students also having used new textbook in Grades 3 and 4
Block C = New textbook - teachers with 1 upgrading
Block E = New textbook - teachers not upgraded
Block B = Old textbook - teachers with 1 upgrading
Block D = Old textbook - teachers not upgraded

From Table 4.1 it can be seen that there are considerable variations between the blocks with regard to pre-test scores. The highest pre-test score is obtained by Block A, and this may be the result of the use over some time of the new textbooks by the pupils as well as the upgrading the teachers received in Grades 3 and 4. This evidence is strengthened by the inter-relation matrix (see Table 4.2), in which the correlation of pre-test with teacher upgrading is highly significant ($r = .247$). In a truly experimental situation such a correlation would not be found, and its presence here indicates the need for an analysis of covariance to obtain unbiased estimates of the textbook and teacher upgrading effects. The gains from pre-test to post-test scores are also presented. The greatest gain (9.81) is obtained by Block B (old textbook, one upgrading only), followed by Block E (new textbook, no upgrading), whereas Block A has the smallest gain.

However, not only were there differences in pre-test scores between the blocks, but similar differences were found in a number of other variables. These included the Raven's Progressive Matrices score, grade repeating, the language used in the home, the language used with friends, learning habits, and the type of lighting in the home. Thus, there were many initial differences between the blocks. It must be remembered that this was a quasi-experimental design and there had been no randomization of students to blocks and of treatment to blocks. Thus, while the gain scores presented in Table 4.1 are unbiased estimates of the students' learning gains during the semester, they cannot serve as a basis for direct comparisons of the effects of textbook or teacher upgrading, because of the initial differences between blocks.

Hence it was necessary to conduct an analysis of covariance to adjust for the initial differences.

ANALYSIS OF COVARIANCE

The basic concept of this analysis is to estimate the effects of independent variables as treatment variables on the dependent variables after adjustment for other independent variables that were considered as covariates. In this study the independent variables which as treatments are the main foci of interest are textbook and teacher upgrading; the dependent variable is student learning as measured by the post-test score in Bahasa Indonesia.

Table 4.2 presents an inter-correlation matrix consisting of the dependent variable (the post-test score in Bahasa Indonesia - variable 1), the treatments (textbook - variable 2, and upgrading - variable 3), together with 16 possible independent variables. Clearly, it would not be appropriate to use all of these as covariates. It was necessary, therefore, to identify a smaller set of independent variables which could together account for the initial differences between blocks.

Table 4.2 Inter-correlation matrix among variables 1/

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Post-test Score	1.00	-.002	.146	.119	.215	.018	-.021	.028	.099	-.006	-.054	.064	-.138	-.007	-.028	-.043	.067	.134	.173
2. Textbook		1.00	.296	.015	.089	.030	-.040	-.008	.000	-.085	-.111	.106	-.011	-.171	-.006	.032	-.042	.020	-.022
3. Upgrading			1.00	.173	.247	.001	.003	.027	-.073	-.089	-.134	.117	-.173	-.048	-.008	-.021	-.169	.063	.027
4. Raven's Progressive Matrices				1.00	.245	-.093	-.068	114	.015	.042	.090	.053	.048	.068	-.057	-.035	-.011	-.066	-.016
5. Pre-test Score					1.00	.020	-.044	.094	.029	.109	.202	-.075	.080	.048	-.010	-.007	-.068	-.100	-.119
6. Sex						1.00	-.081	-.019	-.015	.009	.052	-.037	-.036	.039	.034	.044	-.092	.044	-.019
7. Age							1.00	-.068	-.341	-.046	.076	.011	.043	-.025	-.059	-.041	-.001	-.005	-.199
8. Father's occupation								1.00	-.052	.112	.163	-.044	.051	.187	.086	.068	.086	-.158	-.166
9. No. of grades repeated									1.00	-.002	-.040	.037	-.021	-.023	.010	.006	.044	.053	.214
10. Language at home										1.00	.594	.229	.000	.152	.028	-.032	-.009	-.189	-.201
11. Language with friends											1.00	-.239	.071	.179	.017	-.013	.044	-.222	-.332
12. Illness												1.00	-.040	-.053	.024	.036	.071	.061	.038
13. Learning habits													1.00	.114	.012	-.014	.035	-.095	-.197
14. Type of lighting at home														1.00	.014	.015	-.000	-.219	-.054
15. Parental Status															1.00	.658	.017	-.183	-.050
16. Student living with whom																1.00	.022	-.214	-.026
17. Parental money difficulties																	1.00	-.015	.009
18. Size of family (usually extended) with whom student lives																		1.00	.314
19. Province																			

1/ The scoring scheme for variables is given in Appendix 1.

A systematic elimination of independent variables in Table 4.2 was undertaken. 'Learning habits' which appeared to be a possible candidate as a covariate because of its correlation (-.138) with the criterion, was eliminated for two main reasons:

(1) 'Learning habits' are not necessarily a learning determinant but are part of the learning process.
(2) An examination of the students' answers to this question raised doubts as to whether the less able students had properly understood what was being asked of them.

A number of the other variables were eliminated because of low, non-significant correlations with the criterion. Previous experience had also led the researchers to believe that they would have little effect. Therefore, to decide on a final set of covariates, a small regression analysis was undertaken with the post-test score as the criterion and the following independent variables:

- Pre-test Bahasa Indonesia score;
- Raven's Progressive Matrices score;
- Number of persons living in the students' household;
- Province;
- Sex;
- Age.

The results of this regression analysis are presented in Table 4.3.

Table 4.3 Preliminary regressions using potential covariates on post-test score as the dependent variable

Variables in regression				
Dependent	Independent	R^2	Increment	F
1. Post-test score	Pre-test score	.0483	.0483	50.65 1/
2. Post-test score	Pre-test score + progressive matrices score	.0530	.0046	4.88 2/
3. Post-test score	Pre-test score + progressive matrices score + family no.	.0821	.0292	31.56 1/
4. Post-test score	Pre-test score + progressive matrices score + family no. + province	.1061	.0240	26.7 1/
5. Post-test score	Pre-test score + progressive matrices score + family no. + sex	.1084	.0023	2.65 ns
6. Post-test score	Pre-test score + progressive matrices score + family no. + age	.1097	.0007	1.38 ns

1/ = $p < .01$

2/ = $p. < .05$

ns = non-significant.

Based on these results, the variables of pre-test score, Raven's Progressive Matrices score (as an indicator of student 'ability'), family number at home and province were selected as the covariates. Sex and age added nothing to the analysis and were, therefore, eliminated as covariates.

These variables were then used in further analyses to examine the effect of textbook and teacher upgrading as independent variables on the dependent variable, i.e. the student learning as indicated by the post-test score.

The result of this analysis is presented in Table 4.4.

Table 4.4 Amount of variability on the dependent variable (post-test score) due to the textbook and teacher upgrading

Variables already in the regression	Additional variables	R^2	Increment due to additional variables	F value
0. The covariates	-	.1061	-	-
1. The covariates	1. Textbook	.1062	.0001	.086 ns
2. The covariates	2. Teacher upgrading	.1103	.0042	4.737 1/
3. The covariates + textbook + upgrading	3. Interaction of textbook and teacher upgrading	.1219	.0107	12.11 2/
4. The covariates	4. Textbook + upgrading + interaction	.1219	.0158	5.96 2/

1/ = $p < .05$

2/ = $p < .01$

ns = non-significant.

From Table 4.4 it is apparent that the textbook treatment (either new or old textbook) is most important to the regression when it is considered as an interaction with upgrading (see col. 3). Since the intention of P3D was to improve student learning by supplying both new textbooks and teacher upgrading, it would be expected that the interaction of these would be positive. But we have seen in Table 4.1 that the largest gain was for one upgrading and old textbook. Thus, the gain due to upgrading appears to be smallest for those students using the new textbook. Hence, the upgrading of teachers does not appear to be supporting the changes brought about in the new textbook. As already mentioned, it is to be expected that teacher upgrading is undertaken in order to teach the teachers the new knowledge and skills they will require in order to teach the new curriculum (in this case the new textbook) efficiently. The reader would expect that the upgrading with new textbook would result in better student learning than upgrading with the old textbook. However, the results of the analysis reported above indicate that the teacher upgrading was more effective for the old textbook than for the new one. This casts doubt upon the appropriateness of the teacher upgrading for the new textbook. The observation results reported later in this chapter cast some light upon where the upgrading programme was effective and not effective.

The actual effect of textbook and teacher upgrading on the student learning after adjustment to the covariate variables is presented in Table 4.5.

Table 4.5 Mean scores on pre- and post-text and the effects of adjustment due to the covariates

Mean score	Block study A	B	C	D	E	$\bar{X}$ Total
Pre-test	23.27	17.90	19.03	19.31	17.32	19.31
Post-test (unadjusted)	28.63	27.71	24.98	24.83	25.77	26.46
Post-test (adjusted)	26.67	28.28	25.12	24.89	26.88	-
Adjusted Gain Scores	7.36	8.97	5.81	5.58	7.57	-

THE TEXTBOOK EFFECT

In Table 4.2 we noted that the correlation between post-test score and textbook was -.002. The reader might have thought that this indicated that the type of textbook used had no effect on the final score. However, the covariance revealed a more complicated situation. The textbook effect is seen to be important in its interaction with upgrading. Also in Table 4.5, it can be seen that the adjusted gain score for Block B (one upgrading, old textbook) is greater than for Block C (one upgrading, new textbook) whereas Block E (no upgrading, new textbook) is greater than Block D (no upgrading, old textbook). Therefore, in some circumstances the new textbook is more effective than the old one and in other circumstances the reverse is true. In the item analyses produced later in this chapter the strong and weak points of the two textbooks are shown.

THE EFFECTIVENESS OF TEACHER UPGRADING

The evaluation of the effectiveness of the teacher upgrading programme was based on two main sources of information:

(1) Observation of teaching behaviour in the classroom, and
(2) student learning under teachers with no upgrading, one upgrading and three upgradings.

Observation results

There were 22 items on the observation blank which had to be filled by observers. Of these, 12 were closely related to the content of the upgrading programme.

(1) Is the teaching preparation in accordance with the 'instructional system teaching programme'?

(a) yes (b) no

(2) Does the teacher write instructional objectives clearly?

(a) yes (b) no

(3) Is the teacher more enthusiastic in using teaching aids?
(a) yes (b) no

(4) Is the teacher more open to the questions and critics from the students?
(a) yes (b) no

(5) How is the class discipline?
(a) good (b) medium (c) bad

(6) What is the teacher's medium of instruction?
(a) Bahasa Indonesia (b) mixture

(7) Is the teacher's way of teaching clear enough?
(a) yes (b) no

(8) How are the teaching materials mastered by the teacher?
(a) well (b) badly

(9) Does the teacher give homework?
(a) yes (b) no

(10) How does the teacher evaluate student achievement?
(a) by objective test (b) by oral test

(11) How does the teacher use her teaching time mostly?
(a) for teaching the core subject (b) for others

(12) Are the teaching steps used by the teacher in accordance with those taught in the upgrading?
(a) yes (b) no

The results of the observation are presented by province in Table 4.6.

From the percentages found in Table 4.6, it is clear that:

(1) The amount of upgrading affects teaching behaviour positively - the more teachers have been upgraded, the more they perform as the programme requires them to do. However, the effect of three upgradings (69.5 per cent) is not as much as one might have expected, since 37 per cent of the non-upgraded teachers already performed those behaviours that were introduced in the upgrading course.
(2) The teaching behaviour of the upgraded teachers in Lampung is much better than those in Kuningan.
(3) There is no significant difference in classroom discipline (item 5) and in using time for teaching the core subject (item 11) between the upgraded and non-upgraded teachers.
(4) The upgraded teachers give very little homework to the students (item 9), nor do they use the evaluation methods they were taught during the course (item 10), nor do they use the teaching aids (item 3).
(5) There is a significant difference (Chi-squared) between the amount of upgrading and the demonstration of the teaching behaviour presented in Table 4.6.

Upgrading and student learning

Evaluation on the teacher upgrading based on the student learning product focusses mainly on the gain from pre- to post-test score. Table 4.7 illustrates the amount of gain in student learning according to the amount of teacher upgrading.

Table 4.6 Percentage of teachers conforming to different aims of upgrading course

Item	0 x upgrading		1 x upgrading		3 x upgrading	
	Prov. 1	Prov. 2	Prov. 1	Prov. 2	Prov. 1	Prov. 2
1. Teaching preparation	4	21	83	71	100	100
2. Clarity in writing instructional objectives	4	25	50	33	91	75
3. Using teaching aids	0	4	17	25	33	25
4. Responding positively to the students' questions and criticisms	54	21	83	33	100	58
5. Maintaining classroom discipline	100	100	100	100	100	100
6. Bahasa Indonesia (as opposed to local language) used by teachers	46	79	96	88	83	100
7. Clarity of teaching	79	67	100	88	100	100
8. Mastery of the lesson	67	83	100	92	100	100
9. Giving homework	29	4	8	8	17	0
10. Method of evaluating student achievement	0	0	29	0	17	1
11. Way in which teaching time was used	34	46	33	53	37	44
12. Steps followed in teaching	0	13	67	54	100	100
X	35	39	64	54	73	66
X	37		59		69.5	

Table 4.7 Gain from pre- to post-test score due to amount of teacher upgrading

Mean score	Amount of upgrading		
	3 x	1 x	0 x
Pre-test	23.27	18.47	18.32
Post-test (unadjusted)	28.63	26.35	25.30
Post-test (adjusted)	26.67	26.70	25.89
Adjusted gain score	7.34	7.38	6.58

When separate aspects of Bahasa Indonesia were examined, similar results were found.

From the above results, and those in the covariance analysis, it would seem that upgrading has had an effect on certain teaching behaviours and not on others (the feedback from Table 4.6 being important for those P3D personnel responsible for the teacher upgrading programmes); and that one upgrading is certainly beneficial in terms of student learning, although there is some doubt about the extra benefit of three upgradings.

ITEM ANALYSES

Item analyses were carried out in order to identify poorly achieved objectives and to provide the data for a wrong answer analysis. This information is primarily for the textbook writers, but could also be of use to those responsible for teacher upgrading where factors associated with low achievement are those of teacher behaviour. After reviewing the evidence, the following rules of thumb were established. The appropriate parts of the textbook should be revised if:

(a) a group of items measuring the objectives were answered correctly by fewer than 50 per cent of students;
(b) there is a wide variation (40 per cent or above) between blocks in the percentage of students answering correctly; and
(c) there is a decrease from pre-test to post-test.

Items measuring grammatical structure

There were 19 items that measured the ability to manipulate grammatical structure and which were considered to measure four kinds of specific instructional objectives (s. i. o.) in Bahasa Indonesia:

(1) To change indirect speech to direct speech, including punctuation, or vice-versa (items 34, 35, 46, 47, 48).
(2) To separate a sentence into parts according to the right intonation or rhythm (items 36, 37, 38).
(3) To combine two sentences into one logical sentence (items 39, 40, 41, 42, 43, 44, 45).
(4) Sentence completion (items 49, 50, 51, 52).

The percent correct responses on all items, both for pre-test and post-test are given in Appendix 2.

The highest percent correct response (47 per cent) is obtained by Block B (old textbook - one upgrading), followed by Block A (new textbook - three upgradings) 43 per cent, and then by Block D (no upgrading - old textbook) 40 per cent.

Out of the four specific instructional objectives only groups 3 and 4 are reasonably well mastered (50-70 per cent) by all blocks. This part of the textbook is therefore in urgent need of revision.

An s. i. o. which is very poorly achieved is that of 'changing indirect speech to direct speech, including punctuation' (s. i. o. 1). Item 48, for example, is correctly answered by only 11-13 per cent and there is a decrease from pre-test to post-test, while item 34 is correctly answered by 33-45 per cent of students. For this s. i. o. the textbook is in need of revision and it is suggested that in the teacher upgrading, teachers' mastery and the method(s) of teaching, s. i. o. 1 should be emphasized.

Another s.i.o. which is poorly achieved is that of 'separating a sentence into parts according to the right intonation' (s.i.o. 2). The number of students answering correctly is about 35-36 per cent. Students under the old textbook also achieved slightly better than those under the new one.

Items measuring reading comprehension

There are five items measuring two aspects of reading comprehension: recall (items 6, 7) and comprehension (items 8, 9, 10). On average only 39 per cent of children answered the comprehension items correctly, whereas about 60 per cent answered the recall items correctly.

Items measuring ability in dictation

There are five items in this group. The percentages of students answering these items correctly are given in Appendix 3.

The first three items are fairly well mastered by all blocks. These items deal with the use of capital or small letters. The last two items are poorly mastered (18-29 per cent). These items deal with the correct use of direct speech. Textbook revision on the use of direct speech is called for and it would be desirable that, in the upgrading course, teachers have to master this aspect properly.

Items measuring ability in composition

There are again five items in this group. The percentages of students answering these items correctly are given in Appendix 4.

The number of students answering these items correctly is about 19-27 per cent and no difference is obtained due to the impact of the new textbook, nor by the amount of upgrading. The increase from pre-test to post-test is very small (± 2 per cent). Is it that students were not given the basic teaching when they were in Grades 3 or 4? In any case, revision is necessary.

Items measuring vocabulary

There are 23 items included in this group. Of these, 12 items were fairly well mastered by all blocks. No difference is obtained due to the impact of the new textbook, nor to the amount of upgrading.

The revision of the textbook as far as vocabulary is concerned is not very important, although there are some words not yet mastered (see items 12, 16, 18, 19, 25). These will be quickly mastered by students if they are explained and given enough exercises. No serious revision is needed.

Items measuring ability in spelling

There are five items in this group. They concern the use of capital letters and punctuations. The complete items and responses are given in Appendix 5.

The first three items (53, 54, 55) deal with the use of small or capital letters, whereas the second two items (56, 57) deal with punctuation and direct speech. The new textbook block is slightly better than the old one in using small or capital letters, but since fewer than 45 per cent of students answer the items correctly, the presentation in the textbook and the teaching should be revised.

LEVEL OF MASTERY ACHIEVED

P3D decided that if 60 per cent (or more) of all students answer 70 per cent (or more) of all items correctly, this indicates an acceptable level of mastery of the objectives in the textbook or part of the textbook which has been learned by the students. It would also indicate that only minor revisions would be necessary.

However, in this study, the overall achievement was low - it should be pointed out that the test was not judged to be difficult - and hence it was decided that the level of mastery should be regarded as 'adequate' if 50 per cent or more of the students answered the item correctly. Out of 60 items, only 21 (or 35 per cent) items were mastered.

TEACHERS' BACKGROUND AND SCHOOL FACILITIES

Data were collected on teachers' backgrounds and school facilities to serve as general descriptors for Kuningan and Lampung. These results are not given in detail in this report, but are available at BP3K. Certain selected results are given below which are of direct relevance to the textbook teacher upgrading project.

TEACHERS' BACKGROUND

Some small differences which were not statistically significant between the blocks and between the provinces, were thought unlikely to have contributed to the variations in student achievement. Within the sample, it appeared that teachers using the new textbook were rather younger than teachers using the old textbook. Similarly, upgraded teachers were slightly younger than those who had not been upgraded. Teachers in Lampung were rather older and with more teaching experience than their colleagues in Kuningan. As stated above, however, these differences were not thought to have had a significant effect on the results of the experiment.

SCHOOL FACILITIES

There are no differences between blocks or between old textbook and new textbook schools in terms of school facilities. Most of the school buildings are permanent ones, the teacher-pupil ratio is about 1 to 24, and the size of class varies between 25 and 35. All schools have poor libraries and poor equipment. The health conditions of schools are fairly good. The results for some selected school variables are given in Appendix 6.

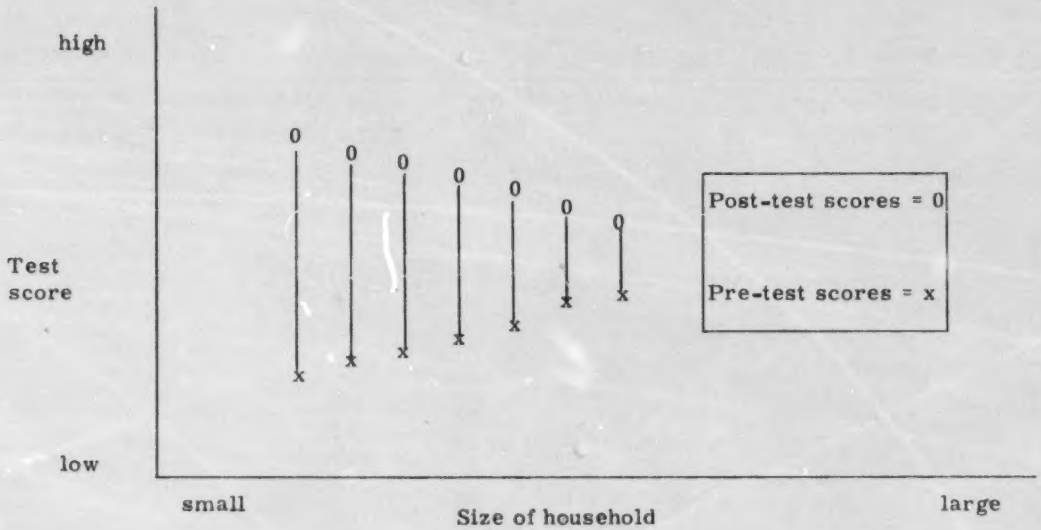
OTHER INTERESTING RESULTS

Number of people in the student's household

Number of people in the student's household was expected to be positively correlated with achievement in Bahasa Indonesia - the larger the family unit, the more the student could learn due to the frequent communication in daily life. In this study it was found that there was a positive correlation (r =.100) as expected with the pre-test score, but with the post-test the correlation was negative (r = -.134).

Thus, the gain in learning Bahasa Indonesia was larger for students from small family groups than from large family groups. Figure 4.1 sketches this phenomenon.

Figure 4.1 Sketch of gain in learning of large and small family groups



The pre-test finding was the same as for the Grade 6 Elementary School Survey. 1/ Further research on this problem is necessary.

Sex

Sex was also expected to be significantly correlated with language achievement. From research in other countries it was expected that at Grade 5 level girls would achieve better than boys. In this study, however, it was found that there were no differences in achievement between boys and girls. The same result was found in the Grade 6 survey already mentioned.

1/ W.B. Elley, National assessment of the quality of Indonesian education, Second Report on the Survey of Achievement in Grade Six (unpublished report), August 1976.

The language spoken with friends and at home was also expected to be correlated with achievement in Bahasa Indonesia. The more students spoke Bahasa Indonesia with friends or at home, the better they would be in Bahasa Indonesia.

In this study students in Lampung spoke Bahasa Indonesia much more than those in Kuningan. However, the post-test scores in Bahasa Indonesia of students in Kuningan were higher than those for Lampung. The explanation for this is not clear and further research would seem desirable.

Subject vs. grade teacher

Of interest to the policy-maker is whether children under subject teachers perform better than those under grade teachers, given the present debate in Indonesia on this matter. The results from this study are inconclusive. A different type of research would have to be mounted to gain firm evidence on this point. Given the importance of this problem, every effort should be made to undertake this research in the near future, but where several subject areas are considered at the same time, not only Bahasa Indonesia.

GENERALIZABILITY OF THE RESULTS OF THIS STUDY FOR OTHER TEXTBOOKS AND TEACHER UPGRADING

The study has been concerned with Bahasa Indonesia only, in the first term of Grade 5 and the results cannot, in any way, be generalized to other subject areas.

However, because of similarities in the procedures used for producing new textbooks in other subject areas and of organising and executing the teacher upgrading, it is possible that certain results of this study may have implications for other subject areas.

CHAPTER 5
GENERAL IMPLICATIONS

NEW TEXTBOOK

The decision has already been taken to produce and distribute new textbooks for all subjects at all grade levels for the transitional curriculum which will remain in force until 1980 or even longer. Thus, the new textbooks will be used whether or not they are better or worse than the old textbooks for effective student learning. The production of new textbooks is a major part of the curriculum development exercise - an exercise which should be conducted systematically, continuously and efficiently.

There are two major reasons why old textbooks should be replaced by new ones: (i) to update the content so that it is more relevant to the present needs of the society as a whole and to the individuals in the society, and (ii) to improve student learning. Thus, the major effort of a curriculum development centre or unit should be devoted to the establishment of a good routine methodology to achieve these aims. This implies, among other things, the building of an appropriate evaluation methodology to the overall work.

The results of the small study undertaken on the first quarter work of Grade 5 Bahasa Indonesia indicated that, on average, the new textbook did not result in better learning than the old textbook, and yet it was reasonable to expect considerably better learning from the new textbook. Hence, it would seem that improvements could be made in the work of the P3D textbook writers unit. It is therefore suggested that:

(i) the methodology of the P3D development procedures be examined and improvements made;
(ii) the textbooks currently being produced be revised where necessary;
(iii) the revision be on the basis of poorly achieved objectives.

OLD VS. NEW TEXTBOOK

Although, on average, there were no differences between the old and new textbooks, there were small differences in particular aspects of Bahasa Indonesia. In particular, the new textbook was slightly worse than the old textbook on grammatical structure, dictation and composition. In more detail the old textbook was better on grammatical structure for:

(a) breaking down sentences according to intonation;
(b) changing direct sentences into indirect ones;
(c) combining two sentences into one logical sentence.

These aspects must clearly be re-examined by P3D. This is from the comparative point of view, but it must not be forgotten that many objectives on both textbooks were poorly achieved and this calls for substantial revision. The formats suggested for revising textbooks are similar to those in the Development School Project.

TEACHER UPGRADING

In general the teacher upgrading had a positive effect on both teaching performance and student learning. In some aspects of teacher behaviour it had very significant effects, but in other aspects not. In detail the important effects were on the writing of teaching preparation, the writing of instructional objectives, responding positively to student questions and criticisms, using Bahasa Indonesia in teaching sessions, clarity of teaching, mastery of the lesson and steps followed in the teaching of a lesson. The amount of upgrading made little or no difference to the use of teaching aids, giving homework and assignments, evaluating student learning and the use of teaching time for the core subject. The failure of the upgraded teacher to perform these last four behaviours in an improved way may well be among the reasons for the small gain in student learning and it is suggested that much more attention be paid to these aspects of the upgrading course.

Efforts should be made to strengthen the upgrading content with respect to the teaching behaviours in:

(a) the use of teaching aids;
(b) the assignment of homework to the students;
(c) the evaluation of student learning; and
(d) the simulation exercises that put more emphasis on using time for teaching the core subject.

This effort should involve intensive supervision of the teachers in the field. Perhaps the most important implication is that much more effort should now go into the formative evaluation of the teacher upgrading courses themselves, and until the programme officers responsible for this work produce a course blueprint, the evaluation work cannot begin.

USE AND EXTENSION OF THE EVALUATIVE PROCEDURES IN FUTURE

Given that curriculum development is a continuing process, the procedures used in this project which might be regarded as more of a formative kind (i. e. the item analyses of the post-test and the possible reasons associated with why students achieve particular objectives poorly, and the analysis of teacher behaviour) should be built into all curriculum development projects. Thus, there would be a high degree of overlap between the evaluation procedures adopted for this type of project and that for the Development School Project.

The five block design used in this project involved a lot of work but provided some interesting results, and in future it is suggested that the comparison and an interaction of treatment variables be undertaken but on a smaller scale.

Thus, if BP3K is to undertake evaluation of other subjects at other grade levels in the P3D project, we suggest the following chronological order:

1. A formative evaluation of the upgrading courses themselves: a test should be constructed (following the blueprint for the course) to be given at the end of the course. Observation will be conducted at set points throughout the course. Both upgraders and those being upgraded will be requested to give specific comments on the difficulties they experience. The item analysis will identify those objectives which are really not being mastered and from the other information collected it should be possible, in nearly all cases, to see what has gone wrong and how it might be improved. The upgraders should then be able to improve their upgrading courses.

2. An evaluation of what the teachers do in their classrooms some time after they have received the upgrading: we suggest that this be done some 6 to 12 months after the upgrading course. Since the objectives of the upgrading course are to improve teaching in the classroom, it is important to see to what extent the teachers are doing what the upgrading course taught them to do. This should be done by observations and comments collected from the teachers on their difficulties and their suggestions for improvements. The analysis would be basically the same as that reported earlier.

3. A formative evaluation of the learning from the textbooks: each textbook can be broken down into a number of sections; this may be by parts of chapter, chapters or sets of chapters. Formative evaluation can be applied to each of these sections and summative evaluation can be made at the end of each semester's work or at the end of the book. The evaluation procedures in this case will be the same as those used in the Development School Project. However, the Wang computer should be used either with mark sense or punch cards to have quick processing of the data.

4. An evaluation comparing either textbooks or amount or types of upgrading: as mentioned above, where it is thought desirable to examine the quality of the textbooks and upgrading, a possible strategy is to conduct a small survey with a pre-test/post-test design. Given the difficulty of randomization for an experimental design and given the high likelihood of contamination in a quasi-experimental design, it is suggested that a survey would be a more realistic approach for this type of evaluation. The survey would include, apart from textbook and upgrading variables, some 20 independent variables, some of which could be used as covariates for determining the 'effect' of the textbook and upgrading variables. The size of the sample would depend on the specific variables being tested, i.e. whether student or teacher variables. Again the unit of analysis would depend on the specific questions being asked, but in general it would be the student as the unit and hence the sample size need not be too large.

If it turned out that randomization could be effected in some two to four groups, then it could be useful to conduct some small-scale experiments, but in general it is very difficult to do this in education.

The above are in chronological order. Priorities are determined by needs and in the immediate future the priority is probably 3, 1, 2, 4.

SECTION III : THE DEVELOPMENT SCHOOL PROJECT

Noehi Nasoetion

CHAPTER 6
DESCRIPTION OF THE DEVELOPMENT SCHOOL PROJECT IN INDONESIA

During the First (1969-74) and Second (1974-79) Five-Year Development Plans, the Ministry of Education has been undertaking and is continuing to conduct various educational reform programmes.

Projects on textbook production, school equipment, school building construction and renovation, pre-service and in-service training of teachers, school administration and organisation are being carried out throughout the country. All of these aim at providing sufficient teachers, curriculum, educational administration and school facilities to sustain the quality, efficiency and relevance of education. Within the framework of administration and curriculum reform one of the projects is the Development School Project (DSP).

The Development School Project is designed to develop a new system of education. In a memorandum of 1970 1/, the Minister of Education and Culture expected that the proposed reform would satisfy the following conditions:

(a) It is advisable that the school represents an integral part of the society around it. According to the principle of life-long education, the school should have two functions. It should be capable of giving both formal and non-formal education for youth and for adult men and women.
(b) It is advisable that the school be 'development-oriented and progress-oriented' so that it can equip the work force with character, knowledge and skill in order to develop the nation in different fields.
(c) It is advisable that the school have a curriculum and a teaching method which is pleasing, challenging and in harmony with that aim.

The above are the basic reasons and objectives for designing an educational reform programme through the Development School Project. 2/ On 2 May 1973, the Minister of Education and Culture stressed the function and the strategies of the DSP. He stated:

1/ Minister of Education and Culture (Minister Mashuri), Basic memorandum on education.
2/ Soedijarto, The modular instructional system as the teaching-learning strategy in the Indonesian Development School, IIEP working paper No. IIEP/S37/6A, February 1976.

"It should be borne in mind that changing an educational system will have many effects on different aspects of life in Indonesia in the future. Therefore, the Development School Project will first of all be conducted as a pilot project using the schools of eight prominent teachers colleges. It is hoped that after sufficient pilot work and after evaluation over a period of time, decisions can be made on its implementation as an educational system. It is in the experimental and pilot stage that we must be careful, accurate, far-sighted, go in depth and be comprehensive. Furthermore, the preparation and planning for the implementation should be well done in order to avoid any negative effects."

THE RESEARCH AND DEVELOPMENT CENTRE (BP3K)

The Chairman of the Research and Development Centre of the Ministry of Education and Culture (BP3K) was chosen as co-ordinator and manager of the DSP. The Chairman of BP3K conducted a series of seminars, workshop and discussion involving educationists, all leaders of the DSP, senior staff of the Ministry of Education and Culture, and leaders of the provincial offices, in order to develop a new programme for the DSP. This is known as the Master Design of Educational Reform to be conducted initially in eight pilot schools.

The head of the Centre for Curriculum Development (a section of BP3K) was assigned as a co-ordinator for the implementation of the master design. He created four working groups which he co-ordinates:

(1) Curriculum development group.
(2) Evaluation group.
(3) Guidance and counselling group.
(4) Administration group.

The members of these groups are mostly people from the eight teachers colleges.

The curriculum development group is composed of subject matter specialists currently teaching at teacher training colleges but who have had experience in teaching of specific subjects in high schools and at teachers colleges, subject matter teachers teaching in school (chosen according to the grade level they are teaching) and professors of the subject areas at university level. Based on a survey of educational needs and objectives 1/ and on the master design 2/, the team developed the DSP curriculum towards the end of 1973.

In the master design four problem areas were identified. These problems were related to:

(1) Quantity of education.
(2) Quality of education.
(3) Relevance of education.
(4) Efficiency of education.

1/ Soedijarto et al, Identifikasi Tujuan-lujuan Pendidikan (Identification of goals, objectives and targets), BP3K, 1972.
2/ Master Design, Pembaharuan Pendidikan melalui Proyek Proyek Perintis Sekolah Pembangunan, BP3K.

Several objectives were generated from these general problem areas 1/:

(1) To develop student centred courses in all subject areas for all age levels.
(2) To develop courses at the school level relevant for those proceeding to employment in rural as well as urban areas and also for those continuing their studies in colleges and universities.
(3) To develop courses and a management system which are sufficiently flexible to cater for students of all interests, needs and potentials.
(4) To develop courses that are sensitive to the environment of the individual and which enhances the national attitudes and national identity of the individual.
(5) To develop effective means of delivering the courses, including improved buildings, facilities, materials and teacher education.
(6) To develop a comprehensive and continuous evaluation system of students and courses so that weaknesses can be corrected and strengths exploited.
(7) To develop guidance and counselling procedures so that each student can be assisted to derive the greatest benefit from the educational opportunities available.

After several discussions it was agreed by the Ministry of Education that the above-mentioned objectives could best be achieved efficiently, effectively and equally by all students if the teaching-learning strategy used was the modular system.

MODULAR SYSTEM OF INSTRUCTION

There were three main reasons for the choice of modules as the medium for the teaching-learning process:

(1) There were very few books that could be used as textbooks relevant to the Basic Course Outline of the DSP.
(2) Not many teachers could make maximum use of the Basic Course Outline in the teaching-learning process in the classroom.
(3) The prevalent teaching-learning process in Indonesia was the teacher delivering his knowledge while the students listened. Given the quality of teachers and the high number of new teachers to be produced in a short time to cope with the expansion of the school system it was felt that the prevailing method of teaching had poor results and in particular, children had little will to learn.

Through the modular system of instruction these shortcomings should, hopefully, be overcome so that quality, relevance and efficiency of education would be achieved. In using the modular system it was anticipated that the role of the teacher would be changed from that of an informant into that of a manager and organiser who is always alert in assisting his students. Things that are not clearly understood or doubted by students could be asked of the teacher for further elaboration. In this way the greater part of the teacher's time is spent in giving guidance both to the individual and to the class.

To improve quality, mastery learning as a strategy was built into the modular system. A student may only begin with the next module if he masters 75 per cent of the objectives in the module on which he is working. The extent to which each student masters the objectives is measured by the

1/ Soedijarto, op. cit.

formative test that accompanies each module. If a student fails to achieve the 75 per cent level he is required to do remedial work, but on the other hand enrichment programmes will be available for high speed learners.

Because of the shortage of resources the enrichment and remedial programmes will be developed later after the main modules have been tried out for every subject matter area and grade level. In the meantime, remedial work consists of re-reading the module or reading another relevant textbook. In Indonesia a module is a small package of a teaching programme, consisting of:

(1) Teachers' manual, which contains guidance about:
   - general and special objectives implied in the module;
   - teaching strategy, including relevant educational technology;
   - list of equipment;
   - evaluation procedures; and
   - vocabulary.

(2) Student activity booklet. This constitutes the heart of the system, containing directives on what should be done by the students to reach the objectives.

(3) Student worksheet. This contains provision for recording key points from the module and the execution of activities. This part is permanently retained by the student.

(4) Answer sheet key for student worksheet items. This is kept by the teacher, to be delivered to the students after finishing their activities in order to check their worksheet work.

(5) Student assessment sheet. This contains test items measuring objectives in the module.

(6) Answer sheet key and marking scheme for the test items.

Experimentation phase

The initial development work required that there should be a small scale try-out of all modules.

Eight experimental schools were chosen. Each of these experimental schools is under the supervision of an IKIP. The implementation of this experiment requires personnel having acquaintance with educational experiments. The IKIPs are the most appropriate institutes as they have a laboratory school.

In the existing school system there are 12 grades. With the introduction of the modules a new system will be tried out with only 11 grades. Because of limited resources the experiment is carried out in several stages. The first stage started with Grades 1, 4 and 9 in four subjects, Bahasa Indonesia, English, mathematics, social sciences and physics.

The try-out started in 1975 and by 1977, it is expected that Grades 1-11 will already be using modules as shown in Table 6. 1.

Table 6. 1 Experimental phase

Year \ Grade	1	2	3	4	5	6	7	8	9	10	11
1975	x	-	-	x	-	x	-	-	x	-	-
1976	x	x	-	x	x	x	x	-	x	x	-
1977	x	x	x	x	x	x	x	x	x	x	x

Other fields of study like handicrafts, sports and religious instruction will be developed later after try-out experience has been obtained from the previously mentioned fields of study. The Basic Course Outline for handicrafts, sports and religious instruction was developed in 1975 and the try-out will be carried out in 1977 in Grades 1, 4, 6 and 9.

Thus, it is hoped that by 1979 all grades in the eight Development Schools will be using the modules[1] for all subject areas.

DEVELOPING THE EVALUATION MODELS

The development school programme is expected to bring about improvement to the present educational system. It is also expected that all the data and information collected during the try-out period will serve as a feedback for the maximum improvement of the programme. For this purpose a task force has been set up and mainly assigned to:

(1) set an evaluation strategy to be implemented in the DSP;
(2) work out an efficient system of collecting relevant data and information on project activities; and
(3) establish a method of data analysis which is suitable for producing relevant feedback to the module writers and decision-makers in the DSP.

The data are collected from the eight schools, which are hundreds of kilometres away from each other; thus, a local evaluation team was set up for each school. The data and information sent by each school will illustrate their differences and similarities so that an overall national picture of those schools can be obtained. At the central government level the National Evaluation Team was created and its membership comprise representatives of the try-out schools and the BP3K staff. The International Institute for Educational Planning, Paris, made its consultants available for this team from December 1974 to September 1976. During the first 18 months the team developed an evaluation model which will be implemented in the evaluation of the DSP and other similar projects in the near future.

In the evaluation, the biggest problem appeared to be the amount of data analysis required. In the first year of the try-out it was estimated that there would be 384 modules to evaluate, comprising four subject areas with twelve modules each at four grade levels. Each year consists of two semesters. If these were to be tried out in eight schools with approximately eighty students (two classes) per grade in each school, the number of students' test results would mount to approximately 245, 760. This would have been a formidable task for analysis, especially if each test consisted of 10-20 questions. Considering the magnitude of the problems, the team and the IIEP agreed to concentrate on the evaluation of Grade 4 science for the purpose of developing an evaluation model for this work. The choice was based on the following: the Grade 4 students' reading ability had reached a level where it would be possible to use group-administered written tests. Secondly,

---

1/ Planning, writing and development of DSP curriculum, BP3K, 1974.

the modular system had not yet been introduced to Grades 2, 3 or 5. Bahasa Indonesia was not selected since it was the subject chosen for the Textbook Teacher Upgrading Project. Mathematics and social sciences, especially civics, have special characteristics and thus were not chosen. Mathematics is a prerequisite subject, and civics, a subject of very great national importance and which is devoted to instilling values and attitudes, needs special measurement techniques. Once the model has been made operational and the results of the evaluation of Grade 4 science modules have been obtained, the same model or a modified version of it will be used for the evaluation of the modules for other subject areas at other grade levels.

The model used

As mentioned above it was decided to concentrate on one subject area at one grade level, namely science at Grade 6. The first year of try-out was to be 1975. In 1975 there were two semesters, the first running from January to June, and the second from July to December. In the first semester there were to be eight modules and, in the second semester, eleven modules. Given that the criterion mastery level had been set, viz. that each student should master 75 per cent of the objectives and that 85 per cent of the students should master any one objective, it was decided that there should be formative evaluation of each module and a summative evaluation of the totality of the modules at the end of the year.

- Formative evaluation

Every effort was to be made to collect as much evidence as possible on each module - evidence which would help the module writers revise the module and the teacher educators improve their teacher training course. The following data, therefore, were to be collected for each module:

(1) Record of time taken by each student to complete the module:
   - reading the text and completing the practical exercise;
   - completing the formative test.
(2) Individual item responses on the formative test.
(3) Total score for each student on formative test.
(4) Information on student background variables (student questionnaire).
(5) Detailed records of types of difficulties (and suggested improvements) encountered by students and teachers (both separately and together).
(6) Time used for (through observation schedules):
   - classroom management;
   - distribution of module and equipment
   - use of blackboard;
   - 'classical' explanation/teaching;
   - individual guidance.
(7) Teachers' assessment of any unintended outcomes.
(8) Comments on the basis of informal discussions.

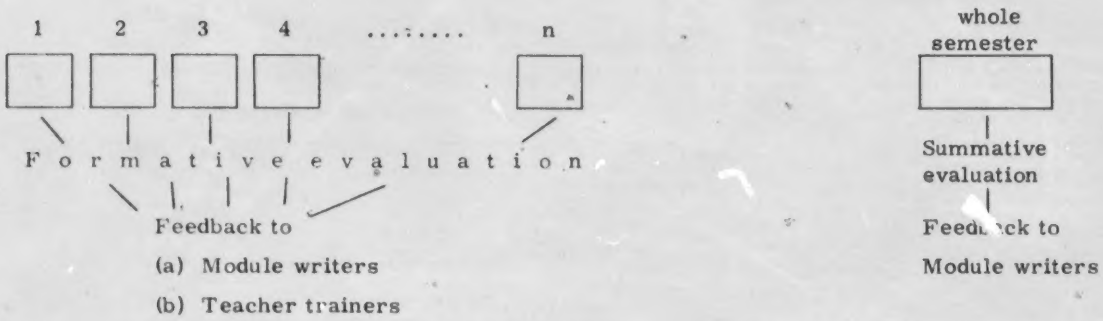
More detail on each of the above is given in Chapter 7. The data collected would be summarized for each school separately and for the eight schools together. Summary sheets of the key findings would be fed back after each module to the module writers and to the teacher trainers. The main emphasis would be on the item analysis of the formative test results and once those objectives are identified which were being poorly achieved, all the other supplementary information would be used for discovering why.

- Summative evaluation

At the end of each semester a summative test would be given. This test would sample the totality of specific objectives covered in the module, ensuring that every general objective was covered. This would allow the strengths and weaknesses of the achievement of the students to be assessed on the general objectives as well as on those specific objectives in the test. Again, the results would be fed to the module writers. At the same time, comments would be made by teachers and students on the set of modules as a whole. Reasons for low mastery of particular objectives would be fed to the module writers after having examined the results of those individual modules in which the particular objectives had been taught and of the comments both on the individual modules and the set of modules as a whole.

Figure 6.1 presents the scheme of the evaluation.

Figure 6.1 Scheme of formative and summative evaluation



A first attempt was made at the evaluation in 1975 and as a result of the experience gained, various aspects of the evaluation were improved to repeat the exercise in Semester 1 of 1976. In 1977 the exercise would include the eight schools included in 1975, but would also be spread to approximately 20 other schools where the contact with the development team would not be so close. In other words, the evaluation would indicate the types of problems which might be encountered in the implementation phase when much of the communication will be restricted to the written word. This model worked well and suggested improvements are made in Chapters 8 and 12.

What follows is a detailed account of what happened in Semesters 1 and 2 of 1975 and Semester 1 of 1976.

CHAPTER 7

FIELD TRIALS IN 1975 AND 1976

INTRODUCTION

Modules are quite new in Indonesia, not only for the people at large but also for education officials. At the beginning only a few knew about the concept of modular instruction. After several discussions, a decision was finally taken by the co-ordinating committee to introduce modular instruction in 1975. In early 1974 the plans went into action: the teachers received in-service training before the end of 1974 and the module writers, the evaluation team and the administrative team made preparations for 1975.

Data to be collected

To collect data as feedback for the module writers, the evaluation team developed, in late 1974, a set of forms for data collection on:

1. Record of time taken by each student to complete each module in terms of: (a) reading text and completing practical exercise; and (b) doing formative test. Analysis of the time spent at different activities in the classroom will give a general impression as to whether the teachers are performing in the way expected of them, or if further instruction or training is necessary. Some summary of average and variation of time required for each module should be made available.
2. Individual item responses on each formative test (in this case each student's answer sheet) were to be returned to Jakarta. The test items are linked to the corresponding instructional objectives in the lesson and worksheet parts of the module. Item analysis would then identify items being poorly achieved by the pupils. Other information would then be examined to discover possible reasons for this poor achievement and to determine if some revision is required in the presentation of the instructional objective in the module - either in the written materials or other aspects of the module presentation.
3. A complete student questionnaire was given, eliciting information on such variables as frequency of grade repeating, student's health condition, size of student's family, conditions of learning at home, parent occupation, etc. This family background information combined with the results of the formative test may be used to determine if some sub-groups of students have more difficulty with or require more time for the module than others. For example, does performance

on the module test increase and time decrease with performance on the formative test; are there differences in performance or time on the modules between boys and girls? Does the development school in Yogyakarta (a rural school) give different results from the other seven schools, and if so, is it because it is rural?

4. Teachers were asked to make a record of:

(a) the page number(s) containing passages where difficulties were encountered by students or teacher for each individual module;
(b) a statement of the type of difficulties encountered in the set of modules (for a semester) by the teacher;
(c) a statement of the type of difficulties encountered in the set of modules by students.

5. Observation schedule. Each school was observed at least once for every module. The observers were members of the local evaluation team. The observation schedules recorded time allocation for:

(a) Classroom arrangement.
(b) Distribution of module and teaching equipment.
(c) Using blackboard.
(d) Classical explanation/teaching.
(e) Individual guidance.

Analysis of the time spent at different activities in the classroom was intended to give a general impression of whether the teachers were performing in the way expected of them, or if further instruction or training was necessary.

Facilities for data processing

As mentioned earlier, there will be an enormous number of data for processing, once evaluation is undertaken for all subject areas at all grade levels. It was considered desirable to have a small computer to cope with this work in order to ensure quick processing of the data. The computer would also be useful to other projects in BP3K.

By the end of 1974, BP3K received a small Wang computer. This computer, employing basic language, was the first of its type to operate in Indonesia and in the first year there was a high frequency of breakdowns. Given the breakdowns and the temporal and financial constraints of using mark sense cards (the Wang computer having a mark sense input device), it was decided to undertake the analysis manually. Basically, only percent right for each item was calculated and such things as a full item analysis (percent right and point biserial correlation for each response), and breakdowns by sex, urban-rural and father's occupation of those not answering an item correctly, would be undertaken in 1977 when the mark sense cards are used and the calculations can be made on the computer.

SEMESTER I, 1975 (FIRST YEAR)

Module try-out and data return

The data collected during the first semester try-out were inadequate. They were incomplete and not all the development schools sent back their completed forms. According to the plans, there should have been eight modules on science for Grade 6, but in practice one of those modules did not reach all schools, i. e. Science Module 06. Hence there were only seven modules tried out. Unfortunately, no one module was ever tried out in all eight development schools. Details of the number of modules tried out, and the schools that tried them out are given in Table 7. 1.

Table 7. 1 Eight modules and development schools that tried them out in Semester I, 1975

DSP	Module No. 01	02	03	04	05	06 1/	07	08	Remarks
1. Jakarta	x	-	-	-	-		-	-	1/ Module 06 did not reach all schools.
2. Bandung	x	x	-	-	-		x	x	
3. Yogyakarta	x	x	x	x	-		x	x	
4. Semarang	-	-	-	-	x		-	-	
5. Malang	-	x	-	x	x		x	-	x = tried out
6. Surabaya	-	x	-	x	x		x	-	
7. U. Pandang	x	x	-	-	-		x	x	- = not tried out
8. Padang	-	-	-	-	-			-	

Six schools had only one class in Grade 6 but in Yogyakarta there were two, in Surabaya three parallel classes.

Reasons for incomplete data

The incomplete data return from the development schools was due to a number of factors such as:

1. Modules did not reach school on time. The reasons for this were:
   (a) The modules were printed behind schedule.
   (b) The manuscripts for modules did not reach the printing press in the right order, leaving the printer to set whichever module reached him first. The setter was not concerned with sequence of modules. As a result module serial numbers could not be administered in their correct order.
2. In the event of a delay in the arrival of a module, the teachers generally taught according to the Basic Course Outline. Some of them worked out their own modules, and only a few would teach according to curriculum developed by the given school instead of relying on the Basic Course Outline.
3. Some difficulties were experienced with the proof-reading of modules being carried out according to schedule. As a result, several printing errors appeared in the modules, illustrations were lacking sharp detail or were incorrectly placed in the module, and in some cases paragraphs or parts of paragraphs were missed out of the module.

4. School administrators did not fully realise the importance of the evaluation part of the try-out for module improvement, with the result that often long delays occurred before the data (often incomplete) were returned to BP3K.

RESULTS

Some observations may be made from the data received and from visits to the development schools. The individualization of the teaching seems to be successful. The teaching-learning process at traditional schools usually calls for the teacher to present the lesson in lecture form and for the students to listen to the lecture, thereby making the teacher a prime source of information. More than 75 per cent of the allotted time is spent on lecturing by the teacher. Under the modular system, on the other hand, the student does much of the reading or the action. The source of information therefore resides in the module, since the teacher comes to the aid of the student only if the latter happens to be experiencing some difficulty. Under the modular system, the teacher spends more than 75 per cent of his allotted time on providing individual or group counselling. However, there are a number of indications that the time for a module is often being cut short, that the percentage of module time spent on the test is too high (as much as 45 per cent), that the estimated time for modules must be increased by a factor of 2 or 3 in many cases, that the score is positively correlated with time taken. The following information was fedback to module writers:

(1) The means and standard deviations (including maximum and minimum) of the time in minutes for all students in each school to complete the text and practical exercise (see Appendix 7).
(2) A summary of the page numbers in each module on which difficulties were encountered (see Appendix 8).
(3) A summary of general difficulties encountered by teachers and/or students (see Appendix 9).
(4) A summary of observations on this new teaching-learning strategy.
(5) Analysis of test relevance which states the objectives of the module, the item measuring each objective, and for each item its content, behaviour (knowledge, understanding or application) and the judged 'validity' of the item to the objective.

From the various data gathered in Semester I, it was determined that this initial phase of the experiment did not run well in terms of the following:

1. The administrative logistics of delivering the various documents to the schools were very poor. Modules did not reach the schools on time and often those that did arrive were not complete. (Each should have contained a Teacher's Guide, a Student's Activity Sheet, a Worksheet, a Test Sheet, a Test Sheet Key and a Marking Scheme).

2. The teachers were not yet used to handling modules - basically a new teaching-learning strategy. The students were also experiencing difficulties. Whereas they had been used to listening to their teachers they now had to read and do things themselves as stipulated in the module they were working on.

3. For module revision purposes the following could be stated:

(a) The flow of evaluation data from the development schools to Jakarta was not smooth.

(b) There were too many evaluation forms to be filled in.
(c) No standard scoring key for the formative test had been produced and teachers had accumulated a total score for students in their school according to their own idea of scoring. The scoring schemes differed from school to school. The results were therefore meaningless for the total evaluation of the modules.

On the basis of the experiences of the evaluation work in Semester 1, the following changes to the evaluation procedures were made:

SUGGESTED CHANGES TO EVALUATION PROCEDURES

1. The information supplied for the module writers by the evaluation team was judged to be too general and not operational enough, e. g. that students had difficulty with a particular term used in module X on page Y. Since it was not explicitly stated, the module writer had to guess which particular term it was meant to be. He expected the term to be mentioned explicitly, in what module it appeared, and on what page. Therefore, the teacher comments request was changed so that the teacher was asked not only to fill in the format on teacher difficulties and student problems, but also to write comments on every word, sentence or page in the module where some difficulty was felt to exist.

2. The development schools' personnel complained about the large number of forms and questionnaires to be filled in at the school, given their limited number of personnel. After a discussion with the local team of evaluators it was decided to discard some forms and improve others. For example, the form relating to students' opinions of the contents of the module was dropped following questionable validity and quality of student responses and the fact that students usually did not complete this form. However, teachers were asked to supply such information after discussions with the students.

3. The summary evaluation reports submitted to the module writers served as a compilation of data covering all the development schools (see Appendixes 7, 8 and 9). Only a few suggestions for improvement of the module were given to the writers. Nevertheless, it was decided that during Semester II efforts would be made to improve such feedback sheets.

4. In order to take advantage of the computer it was decided to purchase some mark sense cards and try out their use. The following action was also taken in preparation for Semester II.

(a) A request was made to the module writers to produce a standard objective marking scheme for each formative test.
(b) The evaluation forms to be filled in by schools were improved, and the local evaluation team (of a given school) received verbal and written explanations on how and what to do at school.
(c) The draft format for mark sense cards was sent by BP3K to the representative of Wang in Jakarta.
(d) BP3K personnel began to prepare a computer programme for item analysis to be available for when the mark sense cards arrived.

SEMESTER II, 1975 (FIRST YEAR)

Semester II started in July. The try-out in this Semester was better than in the first Semester. This was due to the realisation of everyone concerned in the DSP of the weak points experienced in the first Semester. In the second Semester they tried to overcome those weak points.

The module writers were able to finish their drafts earlier, and the evaluation team revised the evaluation forms used by the development schools.

Most of the modules in Semester II were supplemented with marking schemes. Furthermore, it was possible to have a teacher's guide, a student's activity sheet, a student's worksheet, a worksheet key, a test sheet, a test sheet key and marking scheme for most modules. Although not all of these were available for Semester II, at least more were available than in Semester I.

Module try-out and data return

For Semester II there were 11 science modules planned for Grade 6 but only 7 modules were printed and sent out to the schools. All 7 were tried out and an overall view of the try-out is given in Table 7.2.

Table 7.2 Seven modules and development schools that tried them out in Semester II, 1975

DSP	Module No. 09	10	11	12	13	16	17	Remarks
1. Jakarta	x	-	-	-	x	-	-	Modules 14, 15, 18 and 19 did not reach the schools.
2. Bandung	-	x	x	x	x	x	x	
3. Yogyakarta	x	x	x	x	x	x	x	
4. Semarang	-	-	-	-	-	-	-	x = tried out and reported
5. Malang	x	x	x	x	x	x	x	
6. Surabaya	x	x	x	x	x	x	-	
7. U. Pandang	x	x	x	x	-	x	-	- = not tried out and not reported
8. Padang	x	x	x	x	x	-	-	

Unfortunately, many of the experiences of Semester I repeated themselves. Some modules reached schools behind schedule. Four of them had not been fully completed by the writers. When a module did not reach the school, teachers either provided modules of their own, or they proceeded to teach the subject matter under the guidance of the Basic Course Outline.

Comments on Semester II evaluation procedures

The report forms for the module writers were improved and the following information was fed back to module writers on a module feedback form:

1. General results on the time required to complete the module. Try-out result time is computed by first determining the mean value of all eight schools and then multiplying it by a factor of 1.25 to establish the suggested time duration of the module. The value of this factor had been formulated by the team of module writers during a special session with the evaluation team.

2. Teacher difficulties in preparing and using equipment, management of class activities, and administering and scoring tests were reported and categorized.

3. Total test scores for each school and the percentage of students not yet reaching 75 per cent of the objectives for each school were calculated.

4. Comments by evaluators included, among other things:

(a) Test relevance to the specific objectives in each module.
(b) Difficulties encountered by students in the words, sentences and other linguistic aspects used in the module.
(c) Suggestions for the improvement of test items.
(d) A listing of the kinds of activities on which teachers usually spend their time during each module.

A separate feedback form was devised for providing special information on each specific objective not yet mastered by at least 75 per cent of the students. This information is incorporated in a feedback form called Specific Comments on Poorly Achieved Objectives. This form provides information on:

(a) The formulation of the specific objective to be achieved as it appears in the teacher's guide.
(b) The item in the formative test measuring the objective.
(c) Item relevancy (i.e. content validity) to the specific objective.
(d) Percentage of students in each school mastering 75 per cent of the specific objectives.
(e) Teachers' comments on specific objectives and tests as well as on the material appearing on the students activity sheet and students worksheet.
(f) Comments from the evaluation team. These comments are based on the analysis of the data received or on the judgement of the evaluators. These comments were concerned with the formulation of specific objectives, the construction of items and suggestions for other items to be incorporated into the test in future.

Examples of the two feedback forms mentioned above are presented in Appendix 10.

Semester II/1975 saw a total of 7 modules tried out and analysed. These modules contained 53 specific instructional objectives. According to the results of the analysis there were a total of 27, representing 51 per cent of objectives, which the students did not master. The following reasons were offered to explain the non-mastery of particular objectives:

(a) In some cases the items were poor items. Either the stem of the multiple-choice question was presented in a negative sentence form, or the options were of little interest or ambiguous. If these items had been good items, it is estimated that the total mastery would have been nearer to 80 per cent rather than the 49 per cent mentioned above.
(b) In most cases the language used in the module was too difficult for the students to understand. This seems to stem from the use of lengthy sentences and of phrases and terms which were not within the students' normal vocabulary.
(c) Since the range of time taken by students to complete a module was rather narrow, it is assumed that the mastery learning conditions for individualization were not being achieved.
(d) Errors in printing due to the proof-reading not being according to plan.
(e) The formative tests overemphasized the recall aspect of cognitive information and were not in direct accordance with the module blueprints/objectives.

In March 1976, the module writers, especially those dealing with science, stated that the feedback form (see Appendix 10) was appropriate for their needs and requested that this form be used for all future feedback. At the same meeting, a simulation was carried out on the utilization of the evaluation results for Grade 6 science module writers. The information given in that simulation is presented in Appendix 11.

The following examples, taken from Appendix 11, illustrate the type of evaluation feedback produced. This feedback form was found to be particularly useful.

- Formulation of test item was not satisfactory

On the feedback form (see Appendix 11), it is mentioned that the formulation of the "stem of the test item is difficult to understand, and should preferably be written as a positive statement", was the reason for objective 1 not being achieved.

The test item measuring objective 1 was:

Several external factors may effect the crust of the earth.

The factor below that does not effect it is:

(a) The change of the earth at the surface of the crust of the earth.
(b) The change in height at the surface of the crust of the earth.
(c) The flow of water at the surface of the crust of the earth.
(d) The wind that blows over the crust of the earth.

The evaluator suggested the revised test item as: Three main factors that might effect the earth crust:

(a) temperature, water and pressure,
(b) temperature, pressure and air,
(c) water, air and temperature,
(d) air, pressure, water.

- Linguistic problem

On page 7 of the feedback form, teachers commented that the language on the activity sheet about objective 4 was found to be too difficult for the students to understand.

The first activities in the module were:

> You know that the surface of the earth changes from time to time because of magma that comes up from deep inside the earth. But things outside the earth may influence the earth's crust. Such things are temperature, water and air. It is changing temperature which mostly influences the earth's crust. Changes of temperature are associated with changes of season. For example, in our country we have two seasons, i.e. a dry season and a wet season. Night and day temperatures do not vary very much. In other countries, it is quite frequent that the difference of temperature between night and day is very high during the dry season.
>
> You may remember the physical law about the influence of temperature. If something gets hot it expands and if something gets cold it contracts. Thus, what will happen if something gets hot and then gets cold? When this occurs quite often, what will happen to it? Write your answer on your worksheet. Changes like this often occur in areas which have big temperature differences between day and night such as in the Arabian desert, the Gobi desert, etc. Hard stone when it gets hot will expand in the day time but at night the temperature goes down and the stone contracts. Because this occurs over a long period of time, the stone will eventually crack or break.

The evaluator suggested the revised activity:

> As you know the surface of the earth changes from time to time because of magma that comes out from deep inside the earth. Things like temperature, water and air may influence the earth's crust. The change of temperature, because of the change of seasons, is the factor most influencing the earth's crust. In this country we have two seasons, i.e. dry and wet seasons. The temperature does not vary much, either in the dry or wet seasons. In other countries the difference of temperature between day and night is very great.
>
> You may remember the physical law about the influence of temperature on things. Things will expand if they get hot and contract if they get cold. What will happen if a hot thing gets cold very quickly? When this occurs very often, what will happen to that thing? For example, let us look at things in the Arabian desert, the Gobi desert, etc. Hard stone in the desert gets hot and expands during the day, the temperature drops very low during the night, and therefore the stone contracts. Because this occurs very often, the stone cracks or breaks.

According to the team of writers this information is operational enough and is very helpful in efforts towards module revision. The same simulation was extended to all team writers in the four different subject areas. They also favoured the feedback form presented in Appendix 10.

Suggested changes to evaluation procedures at end of Semester II

1. The schedule should be expanded to gather information on the following:
   (a) The extent to which students are actively engaged in learning during the module and the percentage of students who are actively engaged to varying extents.
   (b) The attitude of students and teachers towards the learning strategy embodied in the modular system.
   (c) Whether students completing the module ahead of the other students are given an 'enrichment' assignment, and whether the assignment has any relation to the module.
   (d) Whether there are any 'undesirable' side effects arising from the module, and if so, what they are.
2. The data would continue to be processed manually given no progress on mark sense cards.
3. It was agreed that the observation schedule would be expanded to gather information on student activities, student attitudes, teacher attitudes, 'enrichment' activities and side effects.

As this observation required the services of persons experienced in the teaching-learning process it was suggested that the personnel in charge of the observation would be recruited from subject matter consultants in each development school and from among the staff members of the Committee on Educational Research and Development existing in each school.

Unfortunately the personnel available was limited and hence it was envisaged that only one monthly observation per module per subject area could be made in each school.

General comments from teachers about their difficulties with preparation of equipment, using equipment, management of class activities, administering test and test scoring as found in parts of Appendix 11 were requested.

Summative testing (Semester II)

For Semester II, a summative test was administered at the end of the semester. This test was made by the module writers and in certain aspects they collaborated with the evaluation team. The reason for administering this summative test was to gain experience in constructing and administering summative tests.

SEMESTER I (repeat), 1976

Semester I, 1976 began in mid-January and ended in June. Again, the problems of the administrative logistics required to have the modules revised, printed and delivered to the schools proved to be great. It should be emphasized that the module production and delivery was not the responsibility of the evaluation team but it did mean that no evaluation could be undertaken in those modules not taught.

Module try-out and data return

According to the plan, there were eight modules to be tried out. One of them, Module 6 had not been completed by the module writers. All modules should have reached the schools in the first week of January at the latest. In fact the first modules arrived at Jakarta, Bandung, Yogyakarta and Surabaya in the fourth week of February. Locations farther away from Jakarta only received them in the first week of April. The delay adversely interfered with the flow of evaluation reports to Jakarta. Two development schools, namely Padang and Yogyakarta, were unable to send their evaluation reports because since January they had used the 1975 modules (i.e. unrevised) as a substitute for the revised modules which had not arrived.

Table 7.3 presents a summary of the modules tried out and the data received.

Table 7.3 Seven modules and development schools that tried them out in Semester I, 1976

DSP	Module No. 01	02	03	04	05	06 1/	07	08	Remarks
1. Jakarta	-	-	x	x	x		-	-	1/ Module 06 not
2. Bandung	-	x	x	x	x		-	-	yet put on
3. Yogyakarta	-	-	-	-	-		-	-	manuscript
4. Semarang	x	x	x	x	x		-	-	
5. Malang	-	x	-	x	x		x	x	x = tried out
6. Surabaya	-	-	-	x	x		x	x	
7. Ujung Pandang	x	x	x	x	x		x	-	- = not tried out
8. Padang	-	-	-	-	-		-	-	

Results of Semester I (Repeat), 1976

Because of lack of data on modules (see Table 7.3) only Modules 3, 4 and 5 were analysed. There were 31 specific instructional objectives in the module. Some 61 per cent of the students failed to master at least 75 per cent of those objectives. From the analysis of the accompanying data, the following are believed to be the main reasons for this low level of achievement:

(1) The revisions of the module had not been as good as might have been expected, but this was mainly due to the paucity of feedback data given in Semester I, 1975 (see earlier part of this chapter).
(2) The letters used to caption drawings and illustrations were too small and led to non-understanding.
(3) There were still many words misprinted due to poor or non-existent proof-reading.
(4) Again, about one-fifth of the items were judged to be poorly constructed items. If they had been good only about 40 per cent of the students would have failed to master 75 per cent of the objectives.
(5) Again, mastery conditions of individualization appeared not to be working very well because the range of time required by students to complete a module was very narrow.
(6) In many formative tests there was not a good balance among the behaviours, information, comprehension and application. For example, in one test 80 per cent of items were on recall and in another 90 per cent were on comprehension.

Suggested changes to evaluation procedures

1. The main weakness in the second try-out seems to be the late arrival of modules to the schools, resulting in evaluation data not being able to be collected.

2. The revised observation format reached the development schools in May 1976. The gathering of the observation data began as of that month. Due to limited personnel, time and money, not all of the fields of study and classes tried out were observed, but over the period of one month an observation was conducted on a given field of study in a random class tried out. This observation began from the first part of the module until the completion of the module tried out. The observation format will be continually improved in accordance with comments received from the field.

3. It is not yet possible to carry out data processing with the use of computers, especially the use of mark sense cards to represent student response sheet, because the cards are still on order. It is expected that the cards will be available for application in 1977.

4. The evaluation team and the Director of the whole Development School Project agreed that the Semester II, 1975 would not be revised and that it would be the Semester II, 1976 data which would be used for revision. Firstly, this means that revisions will not be made on the basis of insufficient data (as happened for Semester I, 1975) and that the time pressure will be taken off the module writers.

Summative testing (Semester I, 1976)

A summative test was administered at the end of Semester I, 1976. Generally the results of this test were less beneficial to serve as an input for improvement of the modules. Again, this was not because of poor evaluation but because very few of the modules had been taught in the schools. However, module writers and evaluation personnel gained experience in constructing the summative test together.

CHAPTER 8

FUTURE USE OF DEVELOPMENT SCHOOL EVALUATION PROCEDURES

IN THE DEVELOPMENT SCHOOL PROJECT

The procedures of formative and summative evaluation worked well when there was sufficient data forthcoming from the field. The final form of the feedback formats were much appreciated by the module writers and they found it relatively easy to revise the modules. However, it was suggested by several module writers that they would appreciate it if the evaluators would make direct suggestions for the improvement of certain parts of the modules (e. g. change wording to so and so, change diagram to be like this, etc.). This direct suggestion approach will be experimented with in the coming months by adding a section on to the 'Poorly achieved objective' feedback form.

However, given the difficulties in securing data from all of the trial eight development schools it was suggested that a list of general criteria should be set up and that these criteria should be satisfied before a larger scale trial is initiated. These criteria were:

(1) Approximately 85 per cent of all objectives are mastered by 75 per cent of all students.
(2) There are no deleterious side effects or unexpected outcomes.
(3) The teachers can cope with the system.
(4) The students can cope with the system.
(5) The teachers are 'happy' with the system and their role in it.
(6) The students have (or appear to have) positive attitudes towards the system.
(7) The school principals express satisfaction with the system.
(8) The parents express satisfaction with the system.

The reasoning behind this was that if criteria were not fulfilled for the eight (rather special) IKIP schools, then the probability of the system succeeding in other 'more ordinary' schools would not be very high. On the other hand, there might be more chance of success in ordinary schools than in the IKIP schools. Hence, the plan for 1977 is to add one ordinary school for each of the three levels of the IKIP schools and the 1977 trial will, therefore, be undertaken on 32 schools.

OTHER SUBJECT AREAS AND GRADE LEVELS

Step by step in accordance with existing resources, this evaluation model has begun to be applied in Grade 6 and other grade levels in Bahasa Indonesia, English language, mathematics and social studies. Thus, for example, all teams of module writers have received the feedback form as in Appendix 10 (General Comment on Total Module: I, II, III and IV). The second feedback form (Specific Comment on Poorly Achieved Objectives) will be used to cover all other subject areas, taking the following into consideration:

1. The availability of a specialist for each subject matter to the evaluation team. This will be difficult to achieve in the short run but it is hoped that it will have been fully implemented by the end of 1977. The difficulties of selecting the appropriate subject matter specialist involve:
   (a) The problem of finding a person who is well versed in the given subject matter and in the concepts of evaluation.
   (b) The very limited funds available to warrant additional employment of personnel on the evaluation team.
2. The availability of mark sense cards and the necessary computer programmes for analysing the data. The mark sense cards should arrive in the near future and work will also begin on the programmes. It is hoped that the turn-around time from receipt of mark sense cards and the data analysis can be reduced to two to three days.
3. Some formative tests include open items and in some cases essay questions. This kind of question should be replaced by an objective question (multiple choice, true-false or matching type items) as it facilitates faster and more accurate processing and enables the use of computer facilities.
4. The formulation of very specific objectives in the Teachers' Guide (where only Teachers' Guides exist) should be written in a more precise way. This would make it possible to determine accurately the degree of relevancy between the objectives and the test items.
5. Other subject matters to be tried out for the first time in 1977 are vocational skills, sports and religious instruction. Special measurement techniques will have to be used.

EXTENSION OF MODEL

As can be seen from the preceding chapters the full utilization of the evaluation procedures developed for the project was never realised in that the data for all schools for all modules were never returned - a regrettable fact, but one beyond the control of the evaluation personnel. However, from the evaluation completed on various modules from various schools and in particular, from the Modules 3, 4 and 5 from Semester I repeat, it can be seen that the procedures have worked well.

The data analyses were all done by hand. As long as this was for one subject at one grade level it was manageable but once we progress to several subject areas at various grade levels it will be impossible to undertake the analyses by hand.

Faster analysis

It is, therefore, proposed that mark sense cards be used. The Wang computer owned by BP3K has a mark sense reader attached to it. The reader works and although it currently reads one card per two seconds this time can probably be cut to one card per second. Mark sense cards have been ordered from Singapore and should arrive in the coming months. The children in the schools will have to be trained to complete the cards appropriately, but no particular problem is envisaged. There are plans for the general programmes for the scoring of test results to be produced. The item analyses programmes are operational. Thus, it is hoped that in the very near future it will be possible to process batches of data relatively quickly. The hope is to have a one-week turn-around time from the receipt of data from all (or nearly all) schools to feeding back the information on the poorly achieved objectives to the module writers.

More analysis

Although the test items (or at least most of them) are in multiple-choice form, no full item analysis (including wrong answer analysis) has been undertaken - a task virtually impossible by hand. Once the mark sense card system is working the percent replying to each response for any one item (including omit and not reached), as well as the point-biserial correlation of each response with the total score (minus that item) will be calculated. Furthermore, when fewer than 75 per cent of all students respond correctly, a breakdown of those responding incorrectly (on certain key independent variables such as sex, urban/rural, etc., which the student will record on their mark sense cards) will be calculated.

Other criterion variables

Up to now the evaluation has concentrated on the cognitive outputs of the modular system only. The extent to which children are bored, like the subject, like school and are motivated to do well at school has been judged by observers and teachers. It is desirable to include more objective measures of such affective and motivation outcomes in future and work will begin on the production of these measures as time and resources allow. At the same time it is important to improve the quality of the existing cognitive measures.

Data processing

Many of the basic programmes are now operative on the Wang computer. However, more skill and more persons with skill in the statistical analysis and data processing of educational 'qualitative' data are required. Time must be scheduled on a regular basis for the use of the Wang computer for the data processing of all future module evaluation work.

Personnel

It is also clear that a one-man team will not be able to cope with all of this work and extra personnel will be needed. The present BP3K personnel should provide a combination of formal teaching and learning on-the-job for these support staff.

IN OTHER PROJECTS

Under the Second Five-Year Development Plan, there are still many project activities which remain to be done, particularly with regard to the provision of instructional materials for the primary and secondary school levels. These include:

The Project for Improvement of Primary Schools (P3D)

At the primary school level, the project for basic educational development will in 1977 try out mathematics, science, social studies and Indonesian language at Grade 6 level. Every four months the project gives a so-called post-test to determine to what extent the objectives have been achieved. It is considered that the type of formative evaluation model used in the Development School Project can be used for the basic educational development project.

The Secondary School Textbook Development Project

There are still many student textbooks whose manuscripts are still in a preparatory stage. Should the development of these textbooks include a try-out, then the type of feedback required by the textbook writers will be the same as that required by the module writers in the Development School Project. For this purpose, the team of writers should have set very specific objectives for each textbook they produce. Given this, the evaluators and textbook writers will be able to develop together test items for measuring the objectives. The formative evaluation may be carried out for example twice within one semester namely at the mid-term and end of the semester. The results of this evaluation will serve as feedback to the textbook writers for revision purposes.

The Mass Teacher Upgrading Programme

The Development School Project evaluation procedures can also be used to help improve the teacher upgrading programme carried out on a mass-scale basis both through written courses, radio or television. In this instance, the evaluation forms are to be filled in by those being upgraded and then mailed directly by each person. If all participants were to be tested (some half a million participants) this would create a heavy financial burden. It is therefore suggested that a probability sample be drawn of the target population. The programme of upgrading conducted through written courses, radio or television constitutes a programme which today remains at a planning stage (written courses and television) and at an experimental stage (upgrading through radio). Therefore the evaluation may be carried out without any major difficulty.

SECTION IV : THE EXAMINATIONS (ITEM BANK) PROJECT

Sentot Soelistyo

CHAPTER 9
MODELS FOR NEW EXAMINATION SYSTEMS

After independence when Indonesia began to make determined efforts to provide more and better education for more of its people, the Minister of Education and Culture introduced a 'final state examination' at three final grade levels, namely Grades 6, 9 and 12. This examination was intended to serve two purposes - to act as a 'barometer' to measure the quality of education and to act as a selection device from one stage of education to the next.

In 1968 the Minister decided to abolish this state examination and to institute a school-based examination. Instead of there being one final exam set nationally, each province or sub-province was able to set its own examinations for its own schools. One aim of this change was to delegate more to schools the responsibility for increasing the quality of the education they provide. The period from 1968-72 was spent in developing this school-based examination and 1972 saw the last 'final state examination' and 1973 the first school-based examination.

The national examination was only at the end of one level of education and measured student performance in various subject areas. The 'school-based examination' was also given at the end of each level of education but took into account the teachers' assessments of their students' work during the previous two semesters.

Now a move back to a national examination system is proposed. This has been made necessary by the need to monitor standards in the rapidly expanding public elementary school system with its ever increasing competition for places in secondary and higher education. In a country as large and as diverse as Indonesia, a national examination system must be both versatile and flexible if it is to succeed. The present project aimed to explore some possible models for such a system.

PROBLEMS OF ORGANISING A NATIONAL EXAMINATION SYSTEM

To develop and execute a national examination system for the whole country from the primary up to the senior high school level is not an easy job. There are many factors that could influence the implementation of this system. To communicate with the schools that are scattered all over the archipelago is very hard, especially if one has to visit these schools during the rainy season. Communication facilities to reach them very quickly do not exist. These two factors, the weather and communication difficulties, would be severe problems for the implementation of a national examination system to be carried out over the whole country at one time.

To conduct a national examination also requires a great amount of money. There is the cost of printing the tests and transporting them to all schools. The organisation of the security of the tests is also a problem. Leakage has often occurred in the past in Indonesia and this has required the construction of new replacement tests.

In the past all national examinations have been marked by an examinations committee in each province. The present situation is that the teachers in each school mark the students' examinations from that school. There is no national standard marking scheme. This has resulted in varying standards of marking from school to school. Thus, a score of eight for a student in one school could easily mean something different from a score of eight in another school.

Another problem concerns the advantages and disadvantages of local vs. national norms. If a local norm is used to determine the cut-off point for selection, then those students not being selected locally may be above the national cut-off point if one is used. This might be considered unfair to an average student from a high-scoring area. Conversely, if national norms are used, students of high ability who happen to live in low-scoring areas would be at a disadvantage.

Thus, it is clear that a flexible system of examinations should be produced which can cope with all of the above problems. Since conventional examinations, whether local or national, cannot cope with all of the above problems, other alternative systems were considered and finally it was agreed that experimental work should be undertaken on the item/banking examination system.

ITEM BANKS

The present project has endeavoured to explore the gradual replacement of traditional examinations by a system based on item banks. The concept of an item bank is a recent innovation in education. The idea is that a large pool of test items be classified and calibrated as to their content and difficulty. Within a given subject area, all items are calibrated on the same scale. A test is constructed from the usual blueprint by drawing items from the bank, and scores on the test can eventually be interpreted on the same common scale - regardless of which particular items were chosen.

Advantages of an item banking system compared with traditional examinations

- Flexibility

Traditionally, test instruments have been constructed for use with specific groups of people. The items making up these tests are required to pass certain criteria at a try-out stage when they are administered to a comparable sample of people. In general tests will have different characteristics when applied to new groups of people, and both reliability and validity may change substantially. Furthermore, test scores can usually only be compared where all the students respond to the same set of items. Scores from different tests can only be related where one can make definite assumptions about the distribution of ability in the samples of students who responded to them (e.g. they are identical).

An item bank permits a great increase in flexibility. The statistical indices stored for the items are characteristic of the items themselves and not of the people to be tested. Any test composed of items drawn from the bank will yield scores readily translatable to a fixed scale. Direct comparisons are therefore possible both where different versions of a test are used at the same time, or where individual students are given a series of tests in successive years. The use of different versions of a test may be simply to reduce the amount of copying between students within a classroom, but the concept also extents to using different examinations for students who have studied different curricula, or who live in different regions of the country.

The principle involved is a simple one. Once the items in an item bank have been calibrated, any scores on tests composed of these items will yield consistent estimates of performance on a fixed scale.

- Security

Security is always a problem for tests and examinations. Once they fall into the wrong hands their validity as assessment instruments is placed in doubt.

In Indonesia many people inevitably have access to the tests used in any national examining programme so that control of their security is always a difficulty. In the past it has happened that security has been broken and information about examination questions has been circulated among the teachers and among the pupils prior to the date of the examination. In such a case a new examination needs to be substituted for the one whose security was broken if the results are to be credible, but for carefully constructed tests using conventional methods this is a very time-consuming and expensive process.

With an item bank system the problem of security is considerably less serious. Firstly, access to the entire item bank can be tightly controlled since the bank is centrally located and even if unauthorized people get a sight of several thousand test items it is unlikely that this would ever have a serious effect on examination results. However, if details of a specific test which was to be used in a particular province were leaked, then it would be possible, from the item bank, to replace the test rapidly without any loss of quality and at minimal expense. Lastly, by having available several more or less parallel forms of an examination, and these can of course be easily constructed from an item bank, student cheating and copying on a test can be strongly discouraged.

- Economy

Advantages here stem from the notions of flexibility and security outlined above. The construction and try-out of new items can be reduced as in general it is advantageous to employ the items in an item bank on a number of occasions and in a number of different test contexts. Item validation and calibration is based mainly on past experience of using the bank, and expensive exercises in test development can be avoided once the bank has been created.

- National or local norms

Although an item bank does not itself solve the problem of national versus local norms raised earlier, it does point the way to a solution. When local norms, national norms and an individual's score can all be expressed on the same scale, research can determine the best compromise between the two sets of norms to be used when assessing the individual's level of performance. Cut-off points for selection would thus be based on both sets of norms so as to adjust in an optimal way for regional differences.

REQUIREMENTS OF AN ITEM BANK

The purpose of an item bank is to facilitate the construction of good tests of achievement in a specified subject and of a specified level of difficulty. All items in the bank must be calibrated on a single scale so that results on different items can be readily related. The bank will define a scale of ability/achievement and must have procedures to enable results on tests and examinations drawn from the bank to be expressed on this scale.

Difficulties resulting from the use of traditional item parameters

It should be apparent that it is difficult to meet these requirements using the usual item parameters (facility, discrimination, reliability, validity) unless the bank is only to be used with a very limited population of individuals with known statistical characteristics, such as the ability distribution of that group. In general an item bank will need a set of item parameters that are 'sample-free', that is, that are essentially invariant with regard to the group of people being tested. Such a set is provided by scaling in accordance with a simple mathematical model developed by Rasch. The model and its application to 'sample-free' analysis for item banking is presented in Appendix 12.

Use of Rasch scaling to give a difficulty/ability scale

The sample-free model is applicable to the scaling of homogeneous sets of items (i. e. those measuring the same type of achievement), each set being concerned with achievement in a particular subject. The bank then is divided into a number of sub-banks, one for each subject of study. Each item within a sub-bank is calibrated (i. e. given a parameter describing its difficulty) by Rasch scaling. The relative difficulty of the item is expressed in a unit called a Wit, and the same unit is also used to describe the abilities of people who respond to test items drawn from the sub-bank.

The calibration of items, and also the elimination from the bank of items of poor quality, is accomplished by the analysis of try-out data. For the try-out a set of items is administered to a sample of several hundred people of varying ability. The Rasch analysis and conventional item analysis procedures are thus screens to ensure that only items with good measurement qualities are accepted into the bank and that all such items do exhibit true 'sample-free' behaviour.

IMPLEMENTATION OF THE ITEM BANK CONCEPT

Figure 9. 1 illustrates the entire item banking system. Two main patterns of activity may be noted. The first is concerned with establishing the bank through the design, construction and calibration of new materials. The second with operating the bank, with constructing and scoring tests as a service to users. It should be recognised that each of these activities is repeated within each of the sub-banks that are required.

Procedures for establishing an item bank

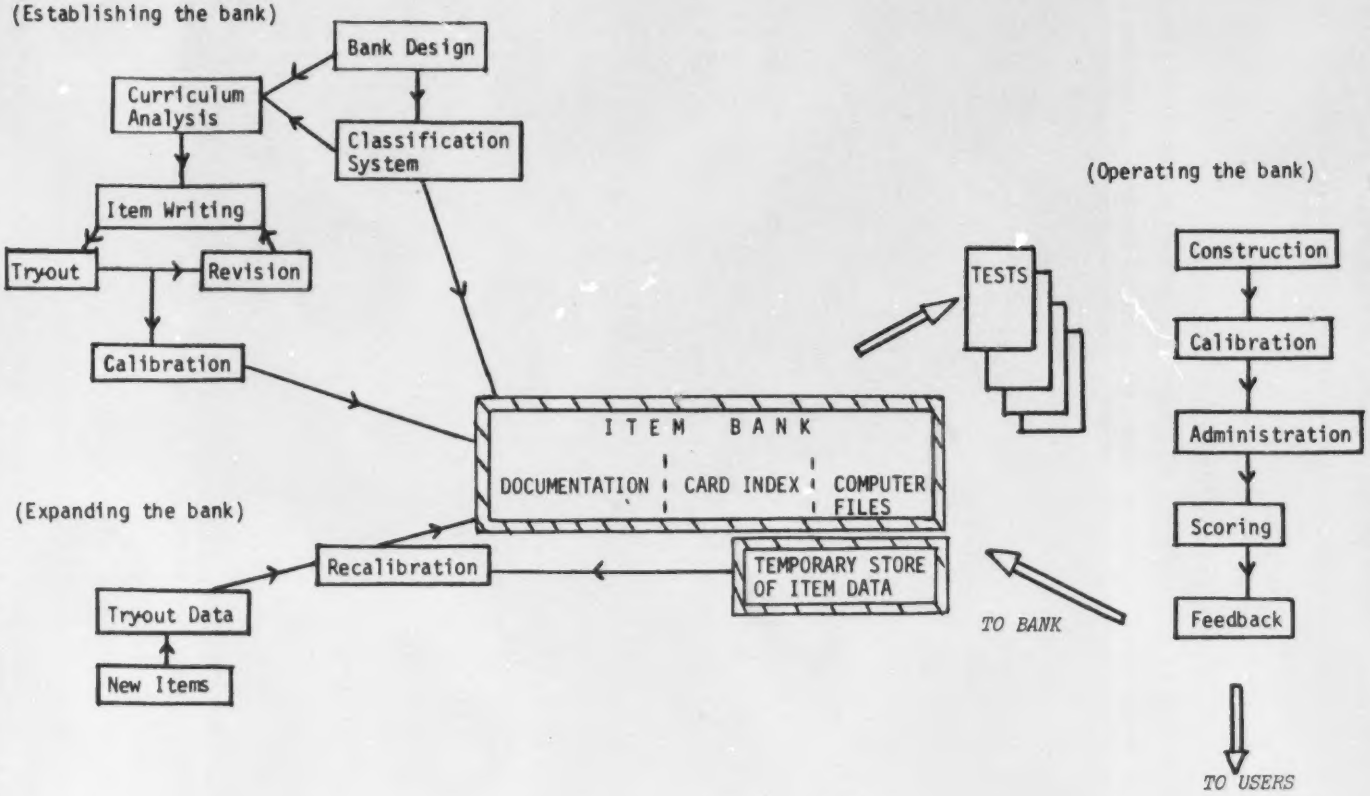
The first stage in creating an item bank will be a thorough analysis of the curriculum in order to compile a list of the instructional objectives whose achievement the bank is intended to assess. A complete classificational scheme for objectives, covering both content and the level of the behaviour required, will be needed. This classificational scheme will also be applied to the test items.

This done, the next step is to commission individuals skilled in the subject matter concerned (e. g. practising teachers) to write items appropriate to the listed objectives. Whilst most experience of item banking work to date has been with multiple-choice items, it should be stressed that Rasch scaling procedures can be similarly applied to any item format that yields unambiguous right/wrong answers. For some subject areas, multiple-choice is probably not the most appropriate format. Thorndike and Hagen (1969) present a useful summary of the evaluation of different test and item types on page 71 of their book. Anticipated use of the bank and financial constraints probably dictate an optimum number of items to be developed. Experience shows that it is sually best to aim at writing twice this number since ambiguous and poorly discriminating items are discarded before the calibration process.

The items are arranged into suitable tests and are piloted with groups of some 400 plus students. The data so gathered is then subjected to two statistical analyses, one a conventional item analysis procedure and the other a special Rasch scaling procedure to check the sample-free nature of the items. The purpose of the first analysis is to identify poorly constructed distractors and other problems caused by the wording of the items. If time permits revisions may be made to some of the items and they may then be included in a second round of pre-testing. The Rasch analysis is specifically to determine whether the items display the required sample-free behaviour with respect to variations in ability. In this analysis the students involved in the pre-test are divided into high and low ability groups on the basis of their overall performance on the trial test. The two sets of data are each used to scale all the items of the trial test. Items that are sample-free should emerge with the same scaled difficulty from both analyses. This behaviour is essential for items that are to be included in the bank. Items that do not have it, usually those with low discrimination, are eliminated from further consideration.

Once the cycle of writing, pre-testing and revising items is complete, the items are calibrated. This will be most easily accomplished if in the trial test forms the items were grouped into overlapping sets. For the calibration it is necessary to be able to estimate the relative difficulty for all pairs of items either directly, if some people were exposed to both items in the pair at the

Figure 9.1 The item banking system



pre-testing stage, or indirectly if each of the two items can be directly related to a third. If this is not possible on the basis of the existing pre-test data then further testing will have to be carried out before the calibration can proceed.

The sample-free scaling by computer, that is described in some detail in Appendix 11, calibrates the relative difficulty of each item in wits.

The resulting set of information forms the nucleus of the item bank. It is stored in two modes, the difficulty calibration data, together with summarised information on performance of the items in the pre-test, are stored in a computer file. At the same time, the text of each item is entered on an index card together with coding indicating the item content, its relevance to particular curricula objectives, its relative difficulty as evidenced by the calibration, etc. is recorded. Fuller details of the storage of bank information and its accompanying documentation is given in Appendix 12.

PROCEDURES FOR OPERATING AN ITEM BANK

Unlike some 'open' item banks, in which all the items and their accompanying data are made generally available to teachers, the item banking system described here will be secure. The texts of the individual items held on the item cards will not be generally distributed. Nevertheless it is very important that those concerned with the examination system should have access to information about the bank and what it can do. It is proposed to achieve this by means of a comprehensive users manual. This will consist of a general section explaining the purposes of the bank and its mode of operation, and a series of subject-specific sections which among other things will give full details of the classification scheme for instructional objectives and bank items in that particular subject. Drafts of part of the users manual are attached as Appendix 13 of this report.

There are however a large group of people who will have to deal with examination results expressed in the new units, but for whom the complexity of the information provided in the comprehensive manual will be inappropriate. For their benefit it is proposed to prepare a small booklet explaining in simple terms the meaning of the wits unit, together with illustrative items and a guide to the interpretation of test scores.

Testing

Test construction using the bank, whether one test or several parallel tests is the aim, follows the usual principles of good test construction. A test blueprint should be drawn up to ensure an adequate sampling of both content and of levels of behaviour.

Items are selected from the bank to match this test blueprint and additional criteria concerning the test difficulty. A Rasch scaled test has two difficulty parameters, one the average wits value of the items concerned, and the other, the width of the test (i.e. the range of difficulty values from the easiest item to the hardest). Past experience and reference to curricular materials usually serve as good guides to choosing the appropriate average level of difficulty. Test width may be more of a problem. For general purposes, a width of approximately 25 to 30 wits has usually

been found adequate, although for specialized testing applications this figure may be varied upwards or downwards. It should be remembered that decreasing the width of a test increases its measurement precision near the average level of difficulty, but greatly reduces measurement accuracy towards the easy and difficult ends of the scale. An attempt should be made to distribute the item difficulties uniformly across the range from easiest to most difficult.

To fill a test blueprint whilst controlling for item content, level of behaviour and level of difficulty, all at once, may seem rather complicated. In fact this procedure is considerably facilitated by the summary information stored on the item cards. With a large well-calibrated item bank it should be possible to obtain tests with characteristics very close to the ideal distributions.

Once an acceptable test has been fabricated, the next step is to calibrate it as a measure of ability. This means calculating estimates of the wits of ability required to produce each of the full range of possible scores on the test. A computer programme is available for this calibration (see Appendix 14).

The test is then ready for printing and distribution. After its administration there are two alaternative procedures for analysing the results. The particular circumstances of the testing situation will determine which is the most appropriate.

Procedure 1 - In this case the raw test data (completed answer sheets or test booklets) are returned to the item bank where they are coded and transferred to punched cards for analysis by computer. The computer then scores the test, transforms the scores into wits, and prints out the results while retaining a summary of the data on item performance for incorporation into the bank. This procedure in general will be appropriate where the number of individuals being tested is not large (say 2,000 or less) and only one form of the test is used.

Procedure 2 - In this case the persons responsible for test administration, perhaps in regional centres, score the test booklets or answer sheets by hand and then transform the scores to wits according to a conversion table supplied with the test. A sub-sample of the answer sheets or booklets (say 1,000 or 2,000) is subsequently returned to the item bank for updating purposes. This procedure will be most appropriate where the number of children being tested is large (e. g. 2,000 or more), where the testing is decentralized (say on a regional basis), or where several different forms of the test have been employed.

Whichever procedure is adopted, after scoring the test booklets, information about the performance of candidates is returned to the school authorities and information about the relative performance on items in the test is returned to the item bank for incorporation into the accumulated store.

Updating the bank

The bank's record of item performance, originally based on pre-test data, is augmented with new information as the items are used. This serves several purposes. Firstly, the addition of data from the operational use of the bank increases the precision of the information concerning particular items and specifically the precision of the difficulty calibration. This will be particularly important when,

because of economic or time constraints, the initial set of pre-test data was fairly restricted in size. Secondly, updating can extend the scope of the bank by the incorporation of new items scaled with, and calibrated against, the old. Thirdly. updating is used to check long-term trends in educational achievement. For example, items in certain content areas may become progressively easier or more difficult over time, and this may give a clue to changes in teaching emphasis within the classroom. Item data for the updating is on punched cards, and a computer programme to carry out the procedure has been prepared (see Appendix 14).

Reference

Thorndike, R. L. and Hagen, E., Measurement and evaluation in psychology and education, 3rd edition, John Wiley and Sons Ltd., New York, 1969.

CHAPTER 10

THE TRY-OUT

SELECTION OF GRADE 11 MATHEMATICS

As an initial step for developing an item bank at BP3K, an analysis using the Rasch model was tried out on a Grade 11 mathematics test.

This mathematics test is a part of the Grade 11 maths-science section battery test. The test covers part of the Grade 11 curriculum in mathematics, specifically that concerned with solid geometry. Since this exercise was carried out specifically for the item bank project, the test was given to a sample of 14th Grade students in 33 senior high schools in Jakarta. These students were in their first month of Grade 12. In all 1,150 students took the test.

Categorization of items behaviour

This mathematics test contains 45 items. They are all four-option multiple-choice questions. The time allowed for the test was 90 minutes. An analysis of the 1968 curriculum, the Development School Curriculum and the old and new textbooks was the basis for the blueprint for the construction of these tests. Items were constructed for the specific instructional objectives in the blueprint. The items were written by a team which consisted of IKIP Bandung staff and senior high school teachers from Bandung. The first try-out was on a small sample of 80 senior high school students in Bandung.

The test items cover the ten topics taught at the senior high school. The number of items for each topic was:

Topics	Number of items
1. Geometric shapes	4
2. Position	5
3. Distance	5
4. Prisms	6
5. Pyramids	6
6. Diameter	4
7. Nets	3
8. Three-angled surfaces (plane)	4
9. Volume	6
10. Understanding of a: point, line, plane (surface)	2

The behavioural objectives measured were recall, comprehension and application (cf. Bloom's Taxonomy).

The number of items measuring each behaviour were:

(1) Recall - 9 items
(2) Comprehension - 23 items
(3) Application - 13 items

Analysis

Once the test scripts were returned from the schools, 1,000 were selected for analysis. The results were transferred to punch cards at the Soedarpo computer centre and were then analysed using programmes supplied by Dr. Choppin. Details of these programmes are given in Appendix 14. Briefly, the analysis consisted of a standard item analysis, a check on the sample-free behaviour of the test items, a Rasch scaling of item difficulty, and Rasch ability estimates for each of the possible scores on the test.

Data summary matrix

For a sample-free scaling, the key data is held in a matrix that shows the relative performance on each pair of items. A portion of this matrix is reproduced in Table 10.1. This shows, for each pair of items i and j, how many people answered i correctly and j incorrectly, and vice-versa.

Table 10.1 Matrix of correct/incorrect answers

		Incorrect answers						
	Item	1	2	3	4	5	.....	45
Correct answers	1	0	180	179	184	119	.....	247
	2	243	0	261	222	155	.....	317
	3	193	152	0	137	112	.....	255
	.	.	.	.	.	.	.....	
	.	.	.	.	.	.	.....	
	45	102	109	96	99	71	.....	0

Table 10.1 shows for example that 243 of the 1,000 students answered item 2 correctly but item 1 incorrectly. Similarly, 180 students answered item 2 incorrectly but got item 1 right. This suggests that item 2 is relatively easier than item 1, a result which proves to be true. In the analysis described in Appendix 12, this is the B matrix, where b_{ij} is the number of individuals responding correctly to item i and incorrectly to item j. Thus for i = 1 and j = 2, b_{ij} would be given the value 180.

Difficulty estimates

The complete B matrix was analysed by programme PAIR to produce a Rasch scaling of ability. Table 10.2 gives the difficulty estimates for some of the items, expressed in wits, together with the standard errors of these estimates and the classical item facility. The higher the wits value for an item, the harder it is, and hence the lower the classical facility.

Table 10.2 Difficulty estimates table

Item	Classical facility	Wits	Error
1	.304	50.714	.564
2	.366	49.195	.554
3	.318	50.457	.579
4	.309	50.664	.577
5	.517	45.655	.568
.	.	.	.
.	.	.	.
.	.	.	.
27	.785	38.474	.751
.	.	.	.
.	.	.	.
44	.028	65.023	1.321
45	.160	55.100	.660

The highest percentage of students, .785, answered item number 27 correctly and the lowest percentage of correct answers, .028, was recorded for item number 44. The difficulty estimates for these, the easiest and hardest items, are 38.474 wits, and 65.023 wits. The associated standard errors are .751 and 1.321. In other words, the difficulty of item 27 can be said with 95 per cent confidence to lie in the range 37.0 to 40.0, while with similar confidence we can state that the difficulty of the hardest item (number 44) lies between 62.4 and 67.6 wits.

Ability estimates

Table 10.3 shows the number of students in the sample who obtained each possible total score and also the estimated ability for each score. No ability estimates are given for students who scored zero or 45 because for such students the evidence available is insufficient (see Appendix 12).

Table 10.3 Relationship of wits ability and score

Score	Sample size	Wits ability	Error (Wits)
0	2	...	...
1	0	28.03	5.27
2	2	31.89	3.86
3	3	34.24	3.24
4	4	35.98	2.87
5	5	37.39	2.62
.	.	.	.
.	.	.	.
.	.	.	.
43	0	68.32	3.93
44	0	72.28	5.49
45	0	...	...

According to Table 10.3 there are 5 students who got a score of 5, and the ability estimate for each of these students is equal to 37.39.

Recalibration of item difficulties for two groups

The total group of students was divided into two groups. One group comprised those scoring higher than the medium (19-45), and the second group those scoring 18 or less. Item difficulties are then recalibrated for each of these groups separately. The results are shown in Table 10.4.

Table 10.4 Item calibration for high and low scoring groups

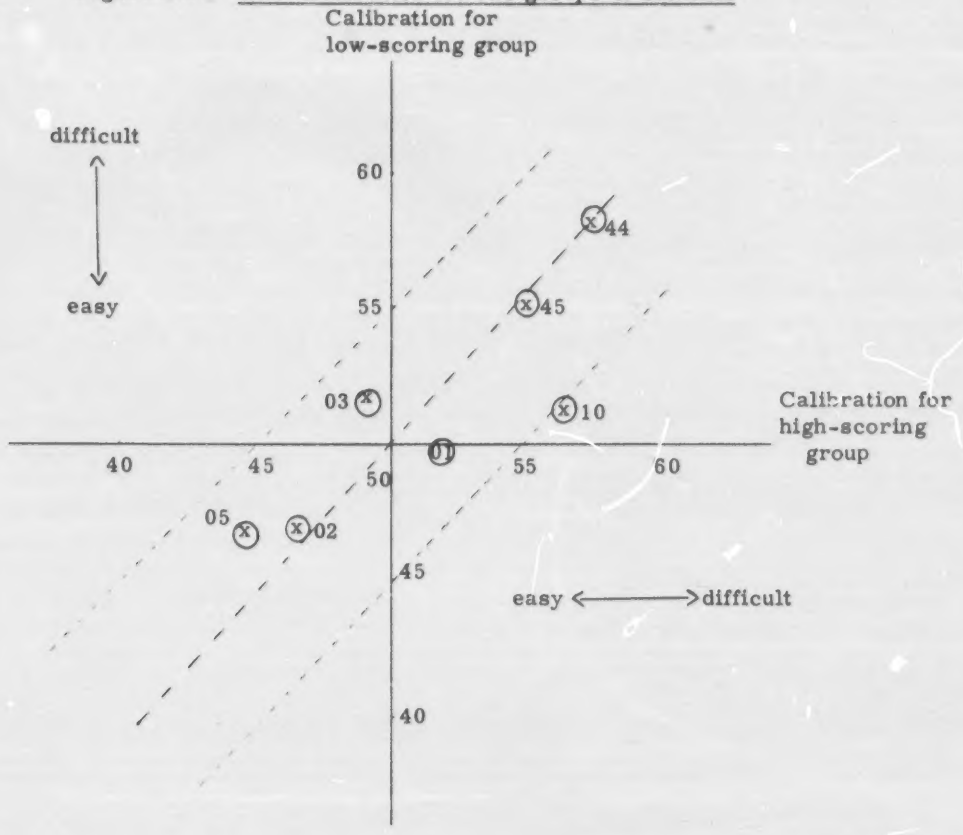
Item	Wits for the low group	Wits for the high group	Standard error (low group)	Standard error (nigh group)
1	49.821	51.257	.920	.720
2	49.189	49.214	.908	.705
3	52.961	49.278	1.044	.718
4	52.470	49.778	1.018	.717
5	47.072	44.691	.873	.761
.	.	.	.	.
.	.	.	.	.
10	52.123	57.087	1.007	.893
.	.	.	.	.
.	.	.	.	.
44	63.845	65.597	2.241	1.639
45	53.534	55.988	1.070	.847

If the items possess sample-free characteristics, one should obtain the same wits estimates from different groups of students. Good items will be those where the two calibrations are the same, within the limits set by consideration of the standard errors.

The discrepancy between the two calibrations is used to decide whether an item should be used or dropped. Usually if between two groups an item has a difference of less than 5 wits, it is considered to be a good item. For example, it can be seen that the difference of wits for item 10 is 4.964 (almost 5) and, hence, it was suggested that item 10 should be dropped from the mathematics test unless it could be improved. Altogether four items from the 45-item test were found not to be sample-free, which is an average proportion considering the limited pre-testing that had been possible.

Figure 10.1 presents the data from Table 10.4 in graphical form. Such a diagram, which can be drawn by computer, makes it easy to see which items fail to satisfy the sample-free criterion. The two axes give the difficulty calibrations of the items in wits, and the dotted lines define the area within which 'sample-free' items should be found.

Figure 10.1 Item calibration for two groups of students



The classical item analysis

Table 10.5 shows the type of computer printout produced. It contains the percentage of right answers of a certain item, response distribution of the four options of each item, the percentage of students who omitted (skipped) the items, and the percentage of students who did not answer the last items in the test (not reached).

Point biserial correlations

Table 10.6 shows the type of computer printout produced. A biserial correlation with the total score for those answering the item correctly is given as the discrimination index. Following this, each column gives the biserial correlation for students choosing a particular option (or 'omitting' or 'not reaching' an item) with the score achieved on a test made up of all items except the one being considered. These correlations give a clear summary of the way the students responded. Thus, for item 1, it is apparent that the highest scoring students chose option 3 (the correct answer). The students who chose option 2 (21.8 per cent according to Table 10.5) were also rather above the average (since the correlation, .113, is positive), and the low ability students chose the other options or did not respond to the item.

Table 10.5 Percent responding to each option on each item

Item	Correct	Omit	Response option 1	2	3	4	N. R.
1	.304	.059	.272	.218	.304 1/	.145	.002
2	.366	.022	.188	.339	.083	.366 1/	.002
3	.318	.091	.318 1/	.164	.210	.215	.002
4	.309	.070	.172	.254	.195	.309 1/	.002
5	.517	.016	.152	.239	.517 1/	.074	.002
.	.	.	.	.	.	.	.
10	.154	.046	.065	.628	.105	.154 1/	.002
.	.	.	.	.	.	.	.
44	.028	.401	.210	.063	.045	.028 1/	.253
45	.160	.000	.201	.160	.160 1/	.177	.301

Note: From the 'not reached' (N. R.) column, it is obvious that there was a substantial percentage of students who did not have time to answer the last items.

1/ right option.

Table 10.6 Point biserial correlations with adjusted total score for each option on each item

Item	Discrimination index	Omit	Response option 1	2	3	4	N. R.
1	.225	-.101	-.086	.113	.147 1/	-.166	-.131
2	.282	-.083	-.031	-.063	-.186	.202 1/	-.131
3	.453	-.107	.380 1/	-.116	-.079	-.200	-.131
4	.395	-.106	-.116	-.042	-.172	.322 1/	-.131
5	.395	-.076	-.152	-.180	.315 1/	-.157	-.131
.	.	.	.	.	.	.	.
10	.065	-.168	-.115	.334	-.133	.001 1/	-.131
.	.	.	.	.	.	.	.
44	.049	.133	-.052	.020	.041	.020 1/	-.034
45	.160	.000	-.098	.020	.097 1/	.113	-.081

1/ right option.

According to Tables 10.5 and 10.6, the first five items appeared satisfactory. Item 10, which was previously identified as failing to meet the sample-free criterion, was not working correctly. Although the correct answer was scored as being 4, no less than 62.8 per cent of the sample selected option 2. Furthermore, the item as scored has almost zero discrimination, and the correlation of .334 for option 2 shows that this was the option chosen by the best students. Further investigation revealed an error in the test scoring since option 2 should have been keyed as the correct answer. With the correct scoring key, this proved to be a successful item which met the sample-free criterion

and could be included in the bank. Item 44, however, which the best students chose to omit, appeared to have very little discriminating power, and was not suitable for further use as it seemed to be too difficult.

The mean, SD and reliability (using the KR 20)

The mean, SD and reliability (using the KR 20) of the test were also shown by the computer printout. For this mathematics test, the Mean = 15, 955; SD = 5, 593; Reliability (KR 20) = . 734.

COMMENTS ON THE TEST

The test seemed to be functioning fairly well, although the large number of omissions towards the end suggested either that the test was too long or that the time allowed was too short. Item 44 was very difficult, perhaps too difficult to contribute much to the measurement.

Some items did not discriminate well and did not display sample-free objective behaviour. They would need to be revised before they could be used in the bank.

The worst items were numbers 10, 13, 19 and 32. Item 10 gave such bad results because of the error in the scoring procedure discussed earlier.

PROBLEMS ENCOUNTERED IN DATA PROCESSING

Several problems were encountered in the data processing of the Grade 11 mathematics test. The major problem was that there was no competent programmer in Fortran on the Research Centre's staff. The classical item analysis programme and the Rasch scaling programme were both written by an IIEP consultant. Some errors in preparing both programmes and data were introduced at the Soedarpo computer centre and this led to difficulties in carrying through the analysis. There had to be continuous and lengthy correspondence with Dr. Choppin in London to resolve these difficulties, and on Dr. Choppin's second visit to Jakarta much time was spent on adjusting the programmes so that after testing they gave the required output. The time taken for processing the data was about four months, but if the mistakes had not occurred the analysis could have been finished in about two weeks. The experience demonstrated the importance of checking carefully each stage of data preparation, particularly the punching of cards, and of the value in having a programming expert on the research centre's staff.

DOCUMENTATION

For an item bank to be used a considerable amount of documentation is required. Basic documentation would include:

A user's manual

(a) objectives of an item bank;
(b) procedures of using the item bank;
(c) classification system;
(d) procedures for constructing the items;
(e) explanation on how to read the computer printouts; and
(f) information on how to use the test booklets and test manuals, which contain the instructional objectives, and the aspects measured by each item.

Item cards

Special cards are needed, one for each item, on which all the relevant data are written. For example, the good items from the Grade 11 test are written on cards. The cards (16 x 22 cm.) are made of manilla paper.

At present there are just 45 item cards for the Grade 11 mathematics test. The card format is given in Figure 10.2.

Since special cupboards to store the item cards are not available yet, the Grade 11 mathematics item cards are put into folders. As the number of items in the item bank increases, so special cupboards and probably a special room or rooms will be needed.

Further details on the documentation required are presented in Appendix 13.

COMPUTER PRINTOUTS AND COMPUTER STORAGE

Besides the manuals and item cards, documentation of the computer printouts is also necessary. The Rasch analysis of the items can easily be read from such printouts.

The data and the computer programmes of the Rasch Scaling are also put onto a disc which is stored in the computer centre.

Reference

Bloom, B.S. et al, Taxonomy of educational objectives, Vol. I: Cognitive domain; Vol. II: Affective domain, New York: David McKay, 1965.

Figure 10.2 Item card format

FRONT

Identification Code 51132.040	Item classification	4 0 Subject	X I Grade Level	0 0 5 Topic or Module

TEXT

If the height of a cut-off Pyramid = 3 dm, the area of the basis = 100 dm^2, and the area of top = 25 dm^2, what is the volume of this cut-off Pyramid?

a. 375 dm^3
b. 175 dm^3
c. 125 dm^3
d. 525 dm^3

Correct answer : ⓑ

SOURCE :

BACK

Identification Code 51132.040	Item classification	4 0	X I	0 0 5

Rasch Scaling Results

Stage of Item	WITS	Std. error	Date
Initial	54.49	65	19-3 1976
Revised			
Revised			
Revised			
Revised			

Use of item

date used	test identification	item number in test	N	REMARKS
23-3 1976	Try-out	40	1000	

Evidence of validity, difficulty, discrimination

Comments on content

CHAPTER 11

FUTURE USE OF THE ITEM BANK IN INDONESIA

The potential is enormous for the use of an item bank system in Indonesia. Everything, however, depends on the rapid development of suitable banks. Listed below are some of the possibilities for the immediate future and longer-term future.

IMMEDIATE FUTURE

SKALU

In the near future, an item bank system will be developed and then enlarged with the items from the entrance examination tests of SKALU (the co-operative secretariats of five universities: University of Indonesia, Institute of Technology, Bandung, Bogor Institute of Agriculture, Gajah Mada, Airlangga University).

The university entrance examination of SKALU is one of the basic tools in the effort to upgrade the quality of higher education as a part of the comprehensive national system of education.

The purpose of the SKALU entrance examination system is to assess the achievement and general knowledge of the candidates so as to predict which will be capable of completing a programme of study at one of the universities or institutes making up SKALU.

It is estimated that about 30,000 will take part in the university entrance examination (SKALU 1977) in the middle of December 1976.

The entrance examination test consists of eight tests:

(1) Biology
(2) Physics
(3) Chemistry
(4) Mathematics
(5) Bahasa Indonesia
(6) English language
(7) Social Studies
(8) Algebra

Each test contains 50 items, and the time allocated for each test is 90 minutes. These items measure all the aspects of Bloom's cognitive domain and the proportions are:

- Knowledge 30%
- Comprehension 40%
- Application 20%
- Analysis, synthesis and evaluation 10%

The items are five-option multiple-choice questions. The types of questions in the test include:

(1) Completion items
(2) Relationship analysis items
(3) Case history analysis items
(4) Multiple completion items
(5) Items using diagrams, pictures and graphs.

For the purpose of the item bank the following exercise will be undertaken. From the 30,000 students a random sub-sample of 2,000 will be drawn and classified into two groups - a high-scoring and a low-scoring group. Analysing these items with the Rasch model will show the wits for each item for each group. Where the difference in wits is < 5 then an item is sample-free and can be used further. On the other hand, where the difference in wits is ≤ 5 the item will be dropped from further analysis. Using the traditional item analysis on the sample of 2,000, weaknesses in the items can be identified and very unsatisfactory items will also have to be excluded. The remaining items will then be subjected to the full Rasch analysis.

The result of the analysis of this 2,000 sub-sample can then be used to scale abilities for the entire sample ($\simeq$ 30,000 students). Raw scores on the items retained for analysis will be calculated by hand using scoring marks. These scores can then be converted to estimates in wits using tables produced by the Rasch scaling analysis. The conversion tables will be similar to that presented in Table 11.1.

Table 11.1 Item wits ability for SKALU tests

Score	Biology Test	Physics Test	Chemistry Test
0	<24	<21	< 25
1	24.3	21.2	25.4
2	25.6	22.3	27.6
3	22.3	24.2	26.3
			and so on

From Table 11.1 it can be seen that a score of 2 on the biology test would imply an estimated ability of 25.6 biology wits while a similar score on the chemistry test would imply an ability of 22.3 chemistry wits.

In the initial application of these techniques the conversion to wits is not of immediate significance. The reason for this is that there will be only one type of biology test being used, and the students who perform best will be readily identifiable whether the conversion to the wits scale is used or not. There are two reasons, however, why the conversion to the wits scale should be carried out. The first is that in subsequent years it may be decided that several different forms

of the biology test would be needed, and then it would be essential to be able to convert the scores to a common scale (the wits scale) in order to be able to compare performances on the different tests. The second is that to resolve the problem regarding the use of national and local norms (described in Chapter 9) it will be necessary to do follow-up studies of candidates selected for study in university to see which of them succeeded or failed, and to relate these results to the background of the student and local and national norms, for all candidates for the SKALU examination. Conversion of raw scores to the fixed wits scale will make it possible to use results from previous years to improve selection in the future. Although the above paragraph has been written with regard to the biology examination, the points made apply with equal force to the other subjects.

Expansion of the item bank - Grades 6 and 9

At the beginning of 1977 this item bank will be enlarged with Grade 6 and Grade 9 items. In 1975 in a sixth grade national assessment survey, tests were given to a sample of 14,000 students throughout Indonesia. The tests were:

(1) Science test	- 60 items
(2) Social Studies test	- 72 items
(3) Mathematics test	- 60 items
(4) Bahasa Indonesia test	- 71 items

Classical item analysis has been undertaken on these tests by the survey team. Later on these tests will be analysed using Rasch scaling. In October 1976, a Grade 9 survey will be conducted. The Grade 9 tests cover five areas of study:

Area of study	Sub-test	No. of items
1. Science	- Physics	50
	- Biology	50
2. Social Studies	- Civics	25
	- History	25
	- Geography	25
	- Economics	25
3. Mathematics	- Algebra	40
	- Geometry	40
4. Bahasa Indonesia		108
5. English language		88
	TOTAL	476 items

The try-out sample will be about 8,400 students, in 200 junior high schools located all over the country.

Possibility of storing items from the curriculum projects

At present a project called 'The Development School Project' (DSP)(see Chapters 6, 7 and 8) is being tried out in eight IKIPs located in Java, Sumatra and Sulawesi. The curriculum of these development schools are different from the other schools in Indonesia. Instruction is carried out through modules.

For further development it is intended to collect items based on different curricula such as that of the development schools. Other testing or research institutions and projects could also store their items in the bank for further development.

LONG-TERM USE

Primary, junior high and senior high school final examinations based on the item banking system

Once the item bank is fully developed and groups of items for all levels of schooling are available, final examinations for all levels of schooling could be conducted based fully on this item bank system.

For this purpose the examination for all schools throughout the country need not be one single test as carried out a few years ago, but several parallel tests could be constructed. The time of the examination need not necessarily be the same in all parts of Indonesia, for the test materials could be different from each other. And because any one test can easily be replaced by a parallel test, there should not be any problem over security. Another advantage is that the examination results of the various schools throughout Indonesia could be easily compared with each other because all scores can be brought onto one national scale.

The immediate preparation to be undertaken if this kind of examination system is going to be carried out is the writing of items for the various subject areas in the final grades of the three levels of schooling. Ideally items for each grade (Grade 6 primary school, Grade 3 junior and Grade 3 senior high school) should be constructed for the 26 provinces of the country. That is 26 x 60 test items = 1,560 items for every subject if each test consists of 60 items. If the above number of items is too large, test for 13 provinces (13 x 60 items = 780 items) could be constructed and for the other 13 provinces items from the first 13 provinces could be exchanged. These items should be tried out on a sample of at most 2,000 students and calibrated by using Rasch scaling.

At this initial stage BP3K, as the co-ordinating body, would provide one or two sets of tests for every province. The printing of the test could be done in the provincial office (Kanwil) in each province. The answer sheets and the keys would be provided by BP3K and the Kanwil could do the scoring of the tests.

Because the results of each test in every grade of these various schools throughout the country could be compared, the item bank system is just the same as the examination system conducted in the years before. Another advantage is that by using the item bank system the budget needed for conducting the final examination is very much reduced.

Possible replacement of the present entrance examinations to junior and senior high schools

An alternative procedure for conducting the entrance examinations could be as follows:

(1) Set up an entrance examination committee in each province. This committee would consist of teachers from the level of secondary schooling for which the students are being selected. The co-ordination of the committee would be the head of the subject matter areas at the regional office of the Department of Education and Culture.
(2) This committee would be responsible for producing the test blueprint for all sub-tests in the examination and would give this blueprint to BP3K.
(3) BP3K would provide each provincial committee with the items (in terms of validity and levels of difficulty) for its blueprint.
(4) Each provincial committee would then work out the format of the test and print and distribute the test booklets (and manuals for test administration and scoring) to the schools.
(5) The schools would administer and score the tests.
(6) A rank order of scores would be made for each province and the ability levels would be known.
(7) Once a pass level has been determined (in terms of the number of places available, etc.) students will be selected.
(8) The answer sheets would be returned to BP3K for further analyses.

The above could be considered as a decentralized provincial system of entrance examinations whereby the responsibility lies with the provincial office, but central help is provided by BP3K. Of course, it would always be possible to have a centralized system whereby the Department of Education and Culture would be responsible for the establishment of the blueprints, the test construction, printing, etc. Whichever the system used, there is still the problem of how many subject areas should be tested - a problem requiring further consideration.

SECTION V : CONCLUSION

T. Neville Postlethwaite and Bruce H. Choppin

CHAPTER 12
GENERAL CONCLUDING REMARKS

The project as a whole must be considered a success. The local staff involved had only very limited experience of evaluation prior to this work. As can be seen from this report a considerable amount of experience has been gained from 'learning on the job'. The projects, admittedly with their limited aims, have been successfully completed. Much has been learned about design, instrument construction and data analysis. The small Wang computer at BP3K now possesses a programme library for the basic univariate, bivariate and multivariate analyses to meet the needs of most evaluation projects in the realm of the quality of education, although it would be desirable to strengthen the expertise in writing new programmes, and in modifying existing ones.

Perhaps most important of all, the advantages - one might say, necessity - of building in evaluation to all endeavours in education, particularly at the planning stage, have been recognised and serious efforts are being made to incorporate evaluation as far as possible into the forthcoming pilot work in many important undertakings.

The primary purpose of this total project was to develop evaluation procedures (models) which would be generalizable to larger educational undertakings in Indonesia. In this sense the procedures developed are of more importance than the results of the small projects undertaken. Each of the small projects here are parts of very large projects. The Textbook Teacher Upgrading Project alone has cost some US$40 million and is intended to provide the materials for the transitional (1975) curriculum up to 1980 for all subject areas in all elementary grades. For this evaluation exercise only the first semester of Grade 5 Bahasa Indonesia was selected. The Development School Project is developing a modular system of instruction based on a new curriculum and again is eventually intended to operate at all grade levels for all subject areas. Grade 6 science was chosen for the purposes of developing the evaluation procedures. Examination procedures are needed not only for monitoring the progress in education at different grade levels, but also for selection purposes from primary to junior high school, from junior high to senior high, and also from senior high school to the university. It was decided to establish an item bank to act as the springboard from which such examinations as are required can be undertaken in as flexible a way as possible. From the item bank, different provinces or even different schools could do different examinations (to fit in with their own curricula), and yet the results of all children can be brought onto one national scale.

Grade 11 mathematics was chosen for testing out the procedures for establishing an item bank. It was not within the scope of this project to test out the running of examinations from it, although the steps to be undertaken in order to do this are described in Chapter 11.

The work accomplished and the procedures developed have been reported in detail in the preceding chapters. Suffice it to mention at this point that the development of procedures is not an easy business. Not only must the procedures be appropriate for the particular problems and the culture but they must fit the level of experience of the persons who are to do the evaluation work and the back-up facilities available.

The development, try-out and small-scale execution of the procedures have been accomplished in 22 months. The procedures selected were those considered to be most important and hence there is room for expansion and improvement. The following are some pointers concerning what might be done.

LEARNING OUTCOMES AND THEIR MEASUREMENT

1. To date, the major emphasis in the projects has been on cognitive learning outcomes. However, schooling has other outcomes of comparable importance. Therefore, work should begin on creating measures for outcomes in the attitudinal, perceptual, socio-emotional and psycho-motor domains. Measures of attitudes and value will be critical in the area of civic education.
2. In foreign language testing, it is important to begin work very soon on constructing measures for listening comprehension, writing and speaking. The construction, administration and scoring of speaking tests is a complicated matter.
3. All formative and summative tests for at least Grades 1 and 2 and probably also for Grade 3 will need to be administered individually. Work should begin immediately on the construction and try-out of such tests.
4. The observation schedules so far produced are rough and ready and have yielded useful data. Nevertheless the observation schedules could be improved, both in terms of the number of student and teacher activities measured and in relating specific observations to either general or specific objectives being taught.
5. Finally, it is suggested that where summative tests are used, not only should a random sample of all the specific objectives taught be taken, but a careful check should be made against all the general objectives and in certain cases new items be constructed to measure the general objectives more directly. It sometimes occurs that the whole is greater than the sum of the parts!

The above remarks concern the formal evaluation undertaken by BP3K, but, of course teachers in their training should always be trained to identify the many cues given by students through their workbooks and homeworks, their experiments, their answers to questions and so on, which indicate where help is needed and where certain parts of textbooks are proving to be difficult for all the children. All of this is evaluation data which must be assessed, interpreted and acted upon by the teacher.

The following suggestions concern the use of the computer facilities over the coming months and years. At present there are four computing facilities being used by BP3K. These are the Wang computer (at BP3K), a Univac computer at Soedarpo and IBM 370 installations at Pertamina and Elnusa. It would seem wise to limit the number of installations used to one for small-scale work (the Wang) and one for large-scale work. The Soedarpo service has proved to be unreliable and rather inaccurate. The Pertamina machine functions primarily as a business computer whereas the Elnusa 370 is primarily for scientific purposes and in this sense may offer a service more suitable for BP3K needs. Hence, we suggest it is for BP3K to choose between Pertamina and Elnusa for the large-scale work that needs to be done, and then to establish a good working relationship with the technical staff of the chosen centre.

There are several points to be made about the Wang computer. The punched card reader is working but there is no facility for punched output. This would greatly facilitate the permanent storage of programmes and data, and would also allow communication with other computers. Hence the card reader should be replaced by a reader/punch. The punched card reader has itself been very much underutilized because there is no key-punch machine. This should be added to the present equipment as soon as possible.

The facility for reading mark sense cards has not yet been utilized although the number of formative evaluation projects which could have fast feedback if they were to use this facility are many. Hence, it would be desirable to begin exploratory work immediately on various formats of mark sense cards generalizable to several projects and on programming the Wang for card reading and scoring.

For the IBM installation the SPSS programme library will cover most needs, and extra programmes (such as Rasch Scaling and classical item analysis) can be obtained from other sources. There is no general programme library for educational work available for the Wang computer and these programmes will have to be written by BP3K staff.

Training in programming and the use of computers is required by BP3K staff. For work on the IBM computer it would be desirable to have at least one and preferably two BP3K persons conversant with Fortran and who will act as the contact men with the IBM machine. Although considerable progress has been made in using the Wang machine in the last year and a half, several BP3K staff should be given training in programming and using the Wang. This training should be given high priority.

The progress in many aspects of evaluation has been impressive in BP3K over the last 22 months. We close by offering our congratulations to the Indonesian authors of this report and express our best wishes for their future work.

APPENDIX 1

CODING AND PUNCHING SCHEMES

IDENTIFICATION DATA AND SCORES

1.	Pre-test and post-test: pre-test = 1; post-test = 2	(col. 1)
2.	Province : Lampung = 1; Kuningan = 2	(col. 2)
3.	Textbook : old = 1; new = 2	(col. 3)
4.	Upgrading : 0 upgrading = 1; 1 upgrading = 2; 3 upgradings = 3	(col. 4)
5.	School number : 01-40	(col. 5-6)
6.	Pupil number : 01-45	(col. 7-8)
7.	Sex : boy = 1; girl = 2	(col. 9)
8.	Age in years : 10-11	(col. 10-11)
9.	Father's occupation : categories 1-3 (3 being high)	(col. 12)
10.	RPM score : 00 - 60	(col. 13-14)
11.	Response on each item : a = 1; b = 2; c = 3; d = 4	(col. 21-80)

STUDENT BACKGROUND DATA

1.	Province : Lampung = 1; Kuningan = 2	(col. 1)
2.	Textbook : old = 1; new = 2	(col. 2)
3.	Upgrading : 0 upgrading = 1; 1 upgrading = 2; 3 upgradings = 3	
4.	School : 01-40	(col. 4-5)
5.	Pupil : 01-45	(col. 6-7)
6.	Sex : boy = 1; girl = 2	(col. 8)
7.	Age : 10-15 years	(col. 9-10)
8.	Father's occupation : Categories 1-3 (3 being high)	(col. 11)
9.	Grade repeating : repeated at least once = 1; no repeating = 2	(col. 12)
10.	Language used at home : local language only = 1; mixture = 2; Bahasa Indonesia only = 3	

11. Language with friends : local language only = 1; mixture = 2; Bahasa Indonesia only = 3

12. Illness : has had serious illness = 1; no serious illness = 2

13. Learning habits : no regular home learning = 1; regular home learning = 2 (col. 16)

14. Lighting in home : oil lamp without chimney = 1; wall lamp = 2; kerosene lamp = 3; electric lamp = 4 (col. 17)

15. Marital status of parents : divorced = 1; both dead = 2; one dead = 3; not divorced = 4 (col. 18)

16. Live with whom : somebody else = 1; with one parent = 2; with both parents = 3 (col. 19)

17. Money difficulties : yes = 1; no = 2 (col. 20)

18. No. of persons in house : over 13 = 1; 10-13 = 2; 7-9 = 3; 4-6 = 4; less than 4 = 5 (col. 21)

19. Reading magazine or other materials in Bahasa Indonesia : no = 1; yes = 2 (col. 22)

APPENDIX 2

SUMMARY OF ITEM PER CENT CHANGES FROM PRE-TEST TO POST-TEST

Items	OLD		NEW			Total
	B	D	A	C	E	
1	47.3–58.4	48.4–65.8	55.9–70.0	43.4–60.0	37.2–60.0	46.3–62.8
2	57.3–67.1	60.3–61.2	61.8–67.1	52.7–53.3	54.8–65.5	57.3–63.0
3	48.8–60.7	52.1–56.1	44.8–36.2	39.8–57.2	42.7–50.6	45.0–53.9
4	26.6–29.0	25.1–26.5	36.4–27.0	27.4–25–0	30.5–32.8	29.2–26.2
5	12.0–23.2	26.5–24.4	29.1–20.2	14.2–30.2	12.1–26.0	18.5–24.9
6	55.2–63.4	53.4–67.8	63.6–74.8	56.6–55.7	50.6–54.8	55.8–63.3
7	45.6–58.9	56.2 –55.1	68.1–62.3	55.8–55.7	43.9–59.0	52.1–58.3
8	9.1–14.6	12.3–13.2	23.2–21.7	20.4–16.8	23.0–21.3	17.6–17.6
9	37.8–52.0	39.3–41.8	46.4–45.4	43.4–46.6	29.7–44.1	39.1–46.2
10	26.6–31.0	29.2–29.0	33.6–30.4	34.1–38.4	33.9–32.0	31.4–32.3
11	38.6–30.6	43.8–44.8	33.6–36.7	42.9–38.9	32.6–33.4	38.3–41.0
12	23.2–29.2	25.1–33.1	26.8–46.3	27.4–40.8	20.5–40.0	24.5–37.9
13	39.4–50.6	37.0–61.2	49.1–62.8	37.2–46.6	31.0–49.3	38.6–55.3
14	35.3–52.5	43.4–39.7	43.2–47.8	41.2–41.8	28.0–39.5	38.0–44.4
15	44.0–60.0	45.7–37.6	58.5–67.1	45.1–54.8	39.7–64.66	45.7–62.7
16	11.6–21.0	15.5–10.7	11.4–20.7	16.4–14.4	10.0–12.5	12.9–16.0
17	44.0–59.	46.1–59.1	66.4–69.0	48.3–48.0	36.4–47.4	46.4–58.8
18	16.2–31.0	22.4–23.4	22.3–43.9	31.0–42.3	28.5–38.1	24.0–35.9
19	15.8–33.3	12.3–17.8	26.8–40.7	26.5–20.6	15.9–33.4	19.4–31.0
20	52.7–73.5	56.2–63.2	66.8–75.8	56.2–71.6	53.6–73.4	56.9–71.8
21	21.6–57.5	33.3–60.7	22.3–69.0	32.7–59.1	22.2–60.4	26.3–61.4
22	21.6–52.5	29.2–54.0	48.6–64.0	26.5–59.1	23.4–66.5	29.6–63.3
23	28.6–61.1	33.3–51.5	42.7–72.9	45.6–72.1	28.5–58.1	35.5–63.3
24	11.6–30.5	16.9–40.3	10.0–64.2	25.2–55.7	8.4–51.6	14.3–48.5
25	19.9–38.3	25.6–34.1	29.5–29.4	32.3–30.2	23.8–33.0	26.1–33.1

Items	OLD		NEW			Total
	B	D	A	C	E	
26.	36.1-50.6	15.3-21.4	24.3-48.7	2[illegible].3-54.3	18.8-48.8	24.0-45.2
27	44.4-52.9	40.2-42.3	46.8-52.1	42.0-48.5	41.0-51.1	42.9-49.6
28.	46.3-64.8	52.3-69.7	60.1-66.1	46.0-69.7	43.9-68.3	51.4-66.7
29.	39.8-65.7	46.1-63.2	55.5-75.8	38.5-56.7	31.4-6099	42.0-64.6
30.	41.5-59.3	45.7-55.6	46.4-61.3	42.0-54.8	29.7-58.1	40.9-58.0
31.	51.0-60.2	48.4-60.2	64.1-66.6	52.7-65.8	54.0-72.5	54.0-65.2
32.	41.1-64.33	41.6-63.8	49.5-46.8	38.9-47.5	39.7-50.6	42.1-55.1
33.	27.3-56.6	29.7-50.5	30.5-26.5	8992-30.7	19.7-45.1	27.2-42.0
34.	21.6-45.6	24.2-33.6	28.6-38.7	29.2-32.2	28.0-40.9	26.3-37.6
35.	22.0-42.9	27.9-22.4	18.8-21.2	29.6-26.4	32.6-36.2	26.1-30.2
36.	21.2-39.2	23.7-30.9	38.5-3019	27.8-33.1	23.4-31.1	25.1-33.1
37.	27.0-42.9	31.1-35.2	26.4-37.6	23.0-30.7	28.5-32.0	27.2-35.8
38.	31.5-42.9	25.1-36.7	33.6-36.7	29.3-34.6	38.9-30.6	31.6-36.4
39.	34.9-57.0	39.7-47.4	37.7-43.4	38.9-36.5	27.2-36.7	35.3-44.3
40.	25.7-29.6	33.8-45.4	33.6-41.5	27.4-31.7	22.6-32.0	28.5-35.9
41	30.7-46.1	26.0-43.31	38.4-97.9	30.1-44.7	24.7-30.2	29.5-48.6
42.	24.1-46.5	32.0-50.0	35.9-52.6	31.0-47.1	25.9-46.0	29.6-48.5
43.	42.3-71.6	42.0-69.3	64.1-80.6	37.6-71.6	37.7-6099	44.5-70.9
44.	32.4-58.4	44.3-62.2	54.1-66.6	32.3-53.3	38.9-44.6	40.3-57.0
45.	23.7-48.4	24.7-41.8	33.6-43.4	27.4-30.2	24.7-38.1	26.7-40.5
46.	8.3-10.5	13.2-17.8	25.0-25.1	20.4-15.3	15.9-15.8	16.4-16.8
47.	21.2-27.8	21.5-16.8	19.5-19.8	21.5-16.8	29.7-25.5	22.8-23.5
48.	21.6-13.2	17.8-13.2	11.8-11.1	19.0-12.9	17.2-13.0	17.6-12.7

Items	OLD		NEW			Total
	B	D	A	C	E	
49.	34.0-63.4	39.3-55.1	54.5-64.7	32.7-50.9	31.8-55.3	38.3-58.1
50.	27.8-52.0	29.2-48.9	43.2-48.7	31.4-37.5	28.0-46.0	31.8-46.7
51.	17.8-42.0	21.5-30.1	26.4-42.9	11.1-26.9	13.0-28.3	17.8-34.2
52.	35.3-66.2	36.8-45.4	42.9-70.5	38.1-62.5	36.0-59.0	38.3-61.0
53.	27.2-47.9	25.1-33.6	45.5-49.2	20.4-42.3	20.5-31.6	27.7-41.1
54.	16.6-40.6	13.8-22.9	33.2-39.6	9.3-18.2	15.1-30.2	17.4-30.5
55.	33.6-46.5	30.6-42.3	51.4-64.2	20.4-38.9	28.0-51.1	32.7-48.8
56.	22.8-49.3	22.8-34.1	40.9-48.3	19.0-33.6	25.1-44.6	25.9-42.2
57.	12.4-12.7	18.7-15.3	22.7-18.8	11.5-18.2	11.3-20.4	15.2-17.1
58.	24.1-23.2	26.0-31.1	30.5-28.0	17.7-33.6	24.3-23.7	24.5-27.9
59.	17.4-27.8	16.9-20.9	25.9-22.7	16.4-19.7	18.8-16.2	19.0-21.5
60.	18.3-25.1	18.7-20.9	21.4-22.2	21.2-23.5	25.5-24.1	21.0-23.2

APPENDIX 3

RESPONSES ON ITEMS MEASURING ABILITY IN DICTATION

Items	Old textbook			New textbook				Total
	B(P1)	D(0)	Av	A(P3B)	C(P1B)	E(B)	Av	pre - post
1	58.4	65.8	62.1	70.0	60.0	60.0	63.3	46.3 - 62.8
2	67.1	61.2	64.15	67.1	53.3	65.5	61.9	57.3 - 63.0
3	60.7	56.1	58.4	36.2	57.2	58.6	50.6	45.0 - 53.9
4	20.0	26.5	23.2	27.0	25.0	32.5	28.1	29.2 - 26.2
5	23.2	24.4	23.8	20.2	30.2	26.0	25.4	18.5 - 24.9
	45.9	46.8	46.3	44.1	45.1	48.5	45.9	

APPENDIX 4

RESPONSES ON ITEMS MEASURING ABILITY IN COMPOSITION

Item	Old textbook			New textbook				Total
	B(U1)	D(0)	Av	A(U3B)	C(U1B)	E(B)	Av	pre - post
58	23.2	31.1	27.1	28.0	33.6	23.7	28.4	24.5 - 27.9
59	27.8	20.9	24.3	22.	19.7	16.2	19.5	19.0 - 21.5
60	25.1	20.9	23.0	.2.2	23.5	24.1	22.8	21.0 - 23.2
	25.4	24.3	24.8	24.3	25.6	21.3	23.6	

APPENDIX 5

RESPONSES ON ITEMS MEASURING SPELLING ABILITY

Item	Old textbook			New textbook				Total
	B(U1)	D(0)	Av	A(U3B)	C(U1B)	E(B)	Av	pre - post
53	47.9	33.6	40.7	49.2	42.3	31.6	41.0	27.7 - 41.1
54	40.6	22.9	31.7	39.6	18.2	30.2	29.3	17.4 - 30.5
55	46.5	42.3	44.4	64.2	38.9	51.1	51.4	32.7 - 48.8
56	49.3	34.1	41.7	48.3	33.6	44.6	42.1	25.9 - 42.2
57	12.7	15.3	14.0	18.8	18.2	20.4	19.1	15.2 - 17.1
	39.4	29.64	34.5	44.0	30.2	35.6	36.6	

APPENDIX 6

MEAN OF SELECTED VARIABLES OF SCHOOL FACILITIES BY BLOCKS AND TOTAL

Variables	Old textbook			New textbook				Total	
	B(U1)	D(0)	Av	A(U3B)	C(U1B)	E(B)	Av		
	X	X		X	X	X		X	SD
1. School building	2.50	2.75	2.63	2.50	2.75	2.75	2.67	2.65	1.66
2. Classroom class ratio	2.13	1.75	1.94	2.13	1.88	1.63	1.88	1.90	0.40
3. Average students	2.75	2.88	2.82	2.75	3.25	2.88	2.96	2.90	0.42
4. Library	1.63	2.00	1.82	1.88	1.88	1.88	1.88	1.85	0.23
5. Distance from sub-districts	4.50	3.38	4.19	3.88	4.88	4.75	4.50	4.28	0.62
6. Room of Grade 5 used	2.38	2.50	2.44	2.63	2.63	2.50	2.59	2.53	0.26
7. Aid and lighting	2.88	3.00	2.94	2.88	3.00	3.00	2.96	2.95	0.10
8. Teaching aid of BII	1.13	1.00	1.07	1.50	1.00	1.00	1.17	1.13	0.28

APPENDIX 7
MEAN AND STANDARD DEVIATIONS OF SCORE AND TIME

Module No.	DSPP	Grade	Score						Time taken												% Time required for	
									Text & worksheet (A)			Formative (B)			Total (C)						A	B
			S_X	$\bar{X}$	N_X	X_{min}		X_{max}	S_A	$\bar{A}$	N_A	S_B	$\bar{B}$	N_B	S_C	$\bar{C}$	N_C	C_{min}	R	C_{max}	$P_A = \frac{A}{C} \times 100\%$	$P_B = \frac{B}{C} \times 100\%$
Sc. 06.1.01.	Yg	VI A	4.89	12.39	33	0.0	19.0	19.0	4.35	302.42	33	6.09	20.54	33	7.80	322.96	33	310	30	340	94	6
	Yg	VI B	4.46	19.48	35	11.0	17.0	28.0	24.12	290.62	35	3.22	17.08	35	23.79	307.71	35	245	82	327	94	6
	Jkt	VI B	9.99	73.23	38	49.4	46.0	95.4	60.43	101.18	38	11.95	23.68	38	59.61	324.86	38	105	365	490	92.70	7.29
	Bdg	VI	16.77	66.68	42	14.0	82.0	96.0	17.72	274.27	40	3.39	21.04	42	61.40	282.30	42	24	-	-	-	-
	Up	VI	14.09	45.02	35	15.0	62.0	77.0	20.95	176.45	35	26.89	26.00	35	6.26	202.45	35	200	36	236	87.15	12.84
Sc. 06.1.02.	Yg	VI A	1.40	4.27	35	2	5.5	7.5	0.33	90.05	35	4.84	12.17	35	4.83	102.22	35	95	15	110	88	12
	Yg	VI B	0.70	5.46	35	4	3	7	2.62	90.66	36	2.54	17.08	36	3.24	107.75	36	105	15	120	84	16
	Up	VI	9.85	33.14	35	7	46	53	6.97	141.97	35	2.06	18.17	35	8.65	160.14	35	115	50	165	89	11
	M	VI	18.39	43.64	37	20	70	90	16.26	127.83	37	3.56	16.81	37	17.49	144.64	37	125	55	180	88	12
	Bdg	VI	22.06	43.34	43	0.0	80.0	80.0	11.46	103.27	43	4.07	20.86	43	12.23	124.13	43	98	82	180	83	17
	S	VI A	14.25	55.22	35	30.0	50.0	80.0	26.83	97.97	35	5.16	14.42	35	27.05	112.40	35	67	93	160	87	13
	S	VI B	13.24	54.72	36	25.0	55.0	80.0	82.85	133.94	36	4.48	18.61	34	84.64	155.14	34	120	510	530	88	12
	S	VI C	13.35	44.27	33	20.0	45.0	65.0	33.80	117.70	34	7.02	15.57	33	31.06	134.42	33	80	100	180	88	12
Sc. 06.1.03.	Yg	VI A	2.55	9.11	34	4.0	12.0	16.0	11.55	145.64	34	2.88	10.67	34	11.77	156.32	34	140	41	181	93	7
Sc. 06.1.04.	Yg	VI A	1.25	4.07	35	1.5	5.5	7.0	3.28	138.04	35	3.63	25.11	35	2.36	164.05	35	160	10	170	85	15
	S	VI A	15.93	49.41	34	20.0	57.5	77.5	30.79	121.34	35	5.86	20.48	35	31.93	141.82	35	100	85	185	86	14
	S	VI B	11.02	43.38	34	25.0	38.0	63.0	23.03	209.25	36	6.74	27.14	35	18.00	237.22	35	210	65	275	89	11
	S	VI C	16.37	41.68	33	21.0	56.5	77.5	24.67	200.47	34	5.29	23.72	33	25.49	224.96	33	183	92	275	89	11
	M	VI	14.72	45.76	38	9	63	72	0.00	125	38	4.89	22.23	38	4.89	147.23	38	140	20	160	85	15
Sc. 06.1.05.	M	VI	3.10	10.67	37	4.0	14.0	18.0	0.82	100.13	37	3.93	21.08	37	3.80	121.21	37	115	10	125	83	17
	Smg	VI	1.56	7.14	39	3.0	6.6	9.6	8.33	86.41	39	6.33	24.23	39	8.88	110.64	39	90	30	120	78	22
Sc. 06.1.06.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sc. 06.1.07.	Yg	VI A	3.45	10.30	33	6.0	13.0	19.0	25.25	170.85	34	13.60	37.54	33	29.52	206.75	33	155	123	278	82	18
	Yg	VI B	3.88	15.11	35	5.0	21.0	26.0	20.66	152.85	35	2.73	24.88	35	21.63	177.74	35	135	70	214	86	14
	Bdg	VI	16.32	33.62	41	7.0	63.0	70.0	3.54	283.95	43	3.12	16.17	41	5.05	300.07	41	277	30	307	94.61	5.38
	Mlg	VI	4.78	11.82	34	4.0	18.0	22.0	11.03	219.42	35	5.06	14.70	34	13.54	234.26	34	215	55	270	93.71	6.28
	Up	VI	16.15	63.79	34	17.0	79.0	96.0	26.44	182.25	34	6.05	58.76	34	26.04	241.01	34	160	103	263	75.61	24.38
Sc. 06.1.08.	S	VIABC	14.20	55.71	103	21.0	70.0	91.0	13.99	297.99	103	3.97	20.56	103	14.53	318.51	103	219	113	331	94	6
	Bdg	VI	4.07	11.62	43	4.0	15.0	19.0	14.04	194.73	38	10.35	32.81	38	18.66	227.55	38	165	85	250	85.57	14.42
	Up	VI	17.39	46.61	34	15.0	65.0	80.0	0.00	205.00	34	5.00	33.52	34	5.00	238.52	34	230	15	245	85.94	14.05
	Yg	VI A	1.40	6.14	34	2.5	6.0	8.5	4.93	203.50	34	2.88	17.23	34	6.57	220.73	34	209	27	236	92.19	7.80
	Yg	VI B	0.95	6.92	34	5.0	3.5	8.5	36.10	196.44	34	4.71	17.38	34	34.38	213.82	34	209	29	239	91.87	8.12

APPENDIX 8
TABULATION FORM

DSPP	FORM	RESPOND

STUDENT DIFFICULTIES
Area of study:
Topic :
Module No. :
Semester :
Grade :

UP	JKT
SBY	BDG
YOGYA	SMG
PDG	MLG

No.	Part of the Module	Type of difficulties	Page 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
		1. Terminology																			
		2. Sentences																			
		3. Language																			
		4. Teaching material (Tools)																			
		5. Filling in work sheet																			
		6. Difficulties in following the sequence as mentioned in module																			
		7.																			

APPENDIX 9
INTERVIEW FORM

ADMINISTRATIVE DIFFICULTIES

Fill by
Local evaluation

DSPP _____

Module : .....

Topic : .....

Class : .....

Time : .....

Try out date .....

No.	Condition / Componen+	Compare to the number of students in one class		Arrival		Arrangement	
		Enough	Not enough	On time	Late	Good	Poor
1.	Amount of module						
2.	Teaching Aids						
3.	Teacher and teacher assistance						
4.	Preparation of module						
5.	Administration of data by						
6.	Class						

APPENDIX 10
GENERAL COMMENTS ON TOTAL MODULE

Table 6. a). General comments on total module

RECAPITULATION OF INFORMATION
ON MODULE EVALUATION

1. Area of study: .
2. Module No. :
3. Topic :
4. Semester/
   Quarterly : ..... year .....
5. Class :
6. Date :

I. TIME

1. Time recommended in module :	minutes
2. Time taken in the 8 PPSP :	minutes

II. TEACHER DIFFICULTIES

Kind of difficulties	No. of teachers who have :		Comments
	difficulty	no difficulty	
1. Preparation of equipment			
2. Using equipment			
3. Management of class activities			
4. Administering test			
5. Scoring test			

III. STUDENT ACHIEVEMENT

1. Average total test score for each school and total:

PPSP	1	2	3	4	5	6	7	8	Total
Average score									

2. Poorly achieved objectives (by item):

Objective classification	Item No.	% of student								Average %	Disregard because poor item (see N below)	Expanded in real version
		Jkt	Bd	Yg	Smg	Sby	Ml	UP	Pd			

For detailed comment on each poorly achieved objective, see following sheets.

IV. EVALUATION TEAM's COMMENT

1. Comment on test item:

Item no.	Relevancy to specific objective	Construction of the test	Expanded in real version

2. Others: (including observation schedule data on module as a whole)

Expanded in real version

..... 197..

Evaluator

( )

b) Specific comments on poorly achieved objective

Subject are:

Grade :

Module nos.:

1	Very specific objective No.									
2.	Item measuring above objective No. of item:									
3.	Comments on validity of item									
4	Percent right in the item in each school	PPSP	1	2	3	4	5	6	7	8
		Item:								
5	Percentage of students in each school mastering 75% of objectives on test	PPSP	1	2	3	4	5	6	7	8
		Item:								
6	Teachers' Comments i). Module text									
	ii). Module work sheet									
	iii). Key for module work sheet									
	iv). Test items									
	v). Scoring key for test									
	vi). Teachers' guide									

7	Comments on observation sheet relevant to this objective
8.	Comments from national evaluation team
9	Comments from informal discussions relevant to this objective

APPENDIX 11
EXAMPLE OF COMPLETED FEEDBACK FORM FOR MODULE 13

a. RECAPITULATION OF INFORMATION
ON MODULE EVALUATION

1.	Area of study	: Science
2.	Module No.	: 06. 1. 13.
3.	Topic	: History of earth II
4.	Semester/Quarterly	: II year 1976
5.	Class	: VI
6.	Date	: September-October 1975

I. Time :

1.	Time recommended in module	: 180 minutes
2.	Time taken in the 8 DSPP	: 200 minutes

II. Teacher Difficulties :

Kind of difficulties	No. of teacher's who have		Comments
	Difficulty	No Difficulty	
1. Preparation of equipment.	-	4	Only 4 DSPP filled the format.
2. Using equipment.	-	4	as above
3. Management of class activities.	-	-	
4. Administering Test	-	-	
5. Scoring Test.	-	4	Only 4 DSPP filled the format.

The number of teaching equipment less than the number of student.

III. Student Achievment :

1. Average total test score for each school and total :

DSPP	1	2	3	4	5	6	7	8
Max score 14	9,0	9,26	8,60	-	9,04	9,61	6,28	9,97

2. Poorly achieved objectives (by item) :

Objective	Item No.	of student								Average %	Discard because poor item
		Jkt.	Bdg	Ygy	Smg	Sby	Mlg	UP	Pdg		
1	1	24,6	19,0	29,0	-	46,0	44,4	12,0	78,3	-	/
3	3	12,3	83,3	20,3	-	26,0	30,6	24,0	32,6	-	/
4	4	84,2	61,9	58,0	-	61,0	66,7	16,0	89,1	-	-
5	5	12,3	9,5	13,0	-	13,0	5,6	24,0	10,9	-	/
9	10	45,6	28,6	42,0	-	47,0	63,9	40,0	54,3	-	-

For detailed comment on each poorly achieved objective, see following sheets.

IV. Evaluation Teams Comment :

1. The language used in the module is somewhat difficult to understand.

2. The answers to the completion of the worksheet may confuse the student when they are checked aganist the Worksheet Key, because there are confusing words in the Worksheet Key, e.g.

Worksheet : (K - 1.6)

In picture 1 the experiment showing a stone ..... and afterwards by ..... may result in the stone being .....

Worksheet Key : (K - 1.6)

In picture 1 the experiment showing a stone EXPOSED TO HEAT and afterwards to COLD may result in the stone being CRACKED OR BROKEN or other answers according to observation.

The words " EXPOSED " and " OR BROKEN, OR OTHER ANSWERS ACCORDING TO OBSERVATION," may confuse the student. Would it not be better to delete the word " EXPOSED " or if necessary to write it in the lead rather than in the Key ?

The phrase " OR BROKEN " may be reworded as follows: " CRACKED " (BROKEN IS ALSO POSSIBLE).
With regard to " OR OTHER ANSWERS ACCORDING TO OBSERVATION ", is there any other results of observation except " CRACKED " or "BROKEN" ?

When there is no other possibility, it is better to delete that phrase.

When there are indeed other possibilities, it is advisable to write then down and not leaving them to the discretion of the students.

3. Test items should preferably be written in positive statements.
4. 85.7 % of the test items in the module measures memory. It is suggested to add more items measuring comprehension, or items dealing with application.
5. There are three test items not measuring one of the Specific Instructional Objectives in the module.
6. More than 50 % of the time allocated was spent for individual or class guidance.
7. Instructions on how to do the test should be written on the Test Sheet.

SPECIFIC COMMENTS ON POORLY ACHIEVED OBJECTIVE

Subject Area : Science
Class : VI
Module No. : 06. 1. 13.

Very specific objective : 1

1. Students were able to mention at least three factors that effect the crust of the earth.

Item measuring above :

2. Several external factors may effect the crust of the earth. The factor below that does not effect it is :
   a. The change of the earth at the surface of the crust of the earth.
   b. The change of height at the surface of the crust of the earth.
   c. The flow of water at the surface of the crust of the earth.
   d. The wind that blow over the crust of the earth.

3. Comments on validity of item :
   It measures memory span, and the questions are relevant to the content of the module.

4. Percent right in the item in each module :

DSPP	1	2	3	4	5	6	7	8
	19,30	35,70	13,04	-	21,0	38,88	0	47,83

5. Percentage of students in each school mastering 75 % of objectives on test item :

DSPP	1	2	3	4	5	6	7	8
	24,60	19,0	29,0	-	46,0	44,40	12,0	78,30

6. Teachers comments :
   6.1. Module text : _____
   6.2. Module worksheet : _____
   6.3. Key for module work-sheet : _____
   6.4. Test items : Instruction on how to do the test should be written on the test sheet.
   6.5. Scoring Key for test : _____
   6.6. Teachers' Guide : _____

7. Comments from national evaluation team : Stem of the test item is difficult to understand, and should **preferably** be written in positive statements.

8. Comments from informal discussions relevant to this objective. : _____

SPECIFIC COMMENTS ON POORLY ACHIEVED OBJECTIVE

Subject area : Science
Class : VI
Module No. : 06.1.13.

Very specific objective : 3

1. Students were able to prove the influence of temperature on certain stones.

2. When a stone is burnt and becomes red hot, and afterwards put into water then the stone will become :
   a. **corrosive**
   b. condensed (or more solid)
   c. porous
   d. a different thing **in structure**

3. The cognitive aspect measured is memory and the questions are relevant to the content of the module. The test does not measure the Specific Instructional Objective (SIO), because the SIO mentions " able to prove " where is the test states " able to mention ".

4. Percent right in the item in each module :

DSPP	1	2	3	4	5	6	7	8
	19,30	35,70	13,04	-	21,00	38,88	0	47,83

5. Percentage of students in each school mastering 75 % of objectives on test item :

DSPP	1	2	3	4	5	6	7	8
	12,30	83,30	20,30	-	26,00	30,60	24,00	32,60

6. Teachers comments :

6.1. Module text : ———

6.2. Module worksheet : Page 1, K.1.b, the last line : after it was exposed to ..... the result is .....

6.3. Key for module worksheet. :

6.4. Test item : Ditto objective No.1 and test No.1.

6.5. Scoring Key for test : ———

6.6. Teachers' Guide : ———

7. Comments from Evaluation Team : The test options should be reconsidered and adjusted to the Worksheet Key.

8. Comments from informal discussions relevant to this objective. : ———

SPECIFIC COMMENTS ON POORLY ACHIEVED OBJECTIVE

Subject Area : Science
Class : VI
Module No. : 06.1.10.

1. Very specific objective : 4
Student were able to prove the influence of temperature on certain stones.

126

2. A desert may result from :
   a). The change of the rainy and the dry season in a certain region.
   b). The change of the mountainous and low lands in a certain region.
   c). The continuous change of the hot and the cold temperature in a certain region.
   d). The thickness of the crust of the earth in a certain region.

3. The aspect measured is memory. The question are relevant to the content of the module.

4. Percent right in the item in each module :

DSPP	1	2	3	4	5	6	7	8
	19,30	35,70	13,04	-	21,00	38,88	0	47,83

5. Percentage of students in each school mastering 75 % of objectives on test item :

DSPP	1	2	3	4	5	6	7	8
	84,20	61,90	38,00	-	61,00	66,70	16,00	89,10

6. Teachers comments :
   6.1. Activity Sheet : A linguistic problem is found on page 3. Please make the language simpler.
   6.2. Module Worksheet : _____
   6.3. Key for Module Worksheet: _____
   6.4. Test Item : _____
   6.5. Scoring Key for test : _____
   6.6. Teachers' Guide : _____

7. Comments from Evaluation Team: Option is longer than stem please do the other way.

8. Comment from informal discussions relevant to this objective. : ———

SPECIFIC COMMENTS ON POORLY ACHIEVED OBJECTIVE

Subject Area : Science
Class : VI
Module No. : 06.1.13.

1. **Very specific objective: 5**
1. The student were able to prove the influence of water on the crust of the earth.

2. The influence of night and day, and rain cause stones :
   a. to crack or break,
   b. to become harder,
   c. to become heat and cold resistant,
   d. to remain **unaffected** by night, day or rain.

3. The aspect measured is memory. The questions are relevant to the content of the module.

4. Percent right in the item in each module :

DSPP	1	2	3	4	5	6	7	8
	19,30	35,70	13,05	-	21,00	38,88	0	47,83

5. Percentage of students in each school mastering 75 % of objectives on test item :

DSPP	1	2	3	4	5	6	7	8
	12,30	9,50	13,00	-	13,00	5,60	24,00	10,90

6. Teachers comments :

6.1. Module text : Please do some revision on material **you used on page3.**

6.2. Module Worksheet : ———

6.3. Key for module Worksheet : _____

6.4. Test Item : _____

6.5. Scoring Key for test : _____

6.6. Teachers' Guide : _____

7. Comments from Evaluation Team : Test : The lead states " influence " whereas the option mentions " not to be influenced ".

Therefore, option (d) does not discriminate (Please,revise) Option (c) is also not well-constructed because the lead mentions " stones ".

8. Comment from informal discussions relevant to this obejctive. : _____

SPECIFIC COMMENTS ON POORLY ACHIEVED OBJECTIVE

Subject Area : Science
Class : VI
Module No. : 06.1.13.

1. **Very specific objective: 9**

1. The students were able to prove that wind causes denudation.

2. Which of the series of events below does not picture the formation of sedimentary stones ?

a. Porous earth ..... denudation..... sediment.
b. Porous earth ..... erosion ..... sediment.
c. Denudation ..... erosion ..... sediment.
d. Rain ..... erosion ..... sediment.

3. It measures memory span. The material asked is relevant to the content of the module.

4. Percent right in the item in each module :

DSPP	1	2	3	4	5	6	7	8
	19,30	35,70	13,05	-	21,00	38,88	0	47,83

5. Percentage of students in each school mastering 75 % of objectives on test item :

DSPP	1	2	3	4	5	6	7	8
	45,60	28,60	42,00	-	47,00	63,90	40,00	54,30

6. Teachers comments :

6.1. Module Text : ____

6.2. Module Worksheet : ____

6.3. Key for module Worksheet : ____

6.4. Test Item : ____

6.5. Scoring Key for Test : ____

6.6. Teachers' Guide : ____

7. Comments from Evaluation Team : Stem of the test item should be preferably be written in positive statement.

8. Comment from informal descussions relevant to this objective. : ____

APPENDIX 12

RASCH SCALING FOR ITEM BANKS

1. The sample-free model

An important characteristic of an 'item' bank is that, just as with a 'money' bank, different units deposited and withdrawn from the bank may be related to one another by means of a well-defined currency. Normal methods of test analysis do not provide this property. A test score is normally only interpretable in terms of the content of the test. Scores on two different tests cannot normally be compared directly, but only through the device of transforming to percentiles of a normal distribution of something similar and this results in the interpretation of one individual's score being dependent on who is included in the sample for the test norming.

Such sample dependence is very limiting for measurement. It would be intolerable if, to compare the lengths of various pieces of string, one had to measure all of them with the same ruler. Intolerable, also, if the length we found for a particular piece of string was dependent upon which other pieces of string had been measured. This, however, is the situation we face with traditional forms of test analysis.

The search for some simple property of test items that would be 'sample-free' led to an examination of the behaviour of items considered two at a time. While the overall proportions of correct responses to either or both items is highly dependent upon the ability of the sample of people involved, it was found that one ratio of results remained relatively constant. If we consider two test items numbered i and j, and use b_{ij} to designate the number of people who respond correctly to item i and incorrectly to item j, with b_{ji} similarly defined; then the ratio $\frac{b_{ij}}{b_{ji}}$ is found not to depend on the nature of the sample of people concerned. This leads us to consider whether a model for item response can be developed so that the ratio $\frac{b_{ij}}{b_{ji}}$ can be estimated by a simple function of item difficulty parameters independent of the characteristics of the people forming the sample.

A simple parameterization would be to take item parameters, say D_i and D_j, such that $\frac{D_j}{D_i}$ would be assymptomatically estimated by the observable ratio $\frac{b_{ij}}{b_{ji}}$. In practice it turns out that this is computationally less convenient than using as difficulty parameters δ_i and δ_j, such that $\frac{b_{ij}}{b_{ji}}$ becomes an estimate of the expression $W^{(\delta_j - \delta_i)}$ where W is an arbitrary constant. (In fact W is fixed at $\sqrt[5]{3}$, i.e. at about 1.2457. The reasons for this become apparent later in the discussion.)

This parameterization corresponds exactly to the model for item response:

Probability (person u answers item i correctly)

$$= \frac{W^{\alpha_u}}{W^{\alpha_u} + W^{\delta_i}}$$

where $_u$ is a parameter for the ability of person u. This, which is more often written as:

$$P\left[a_{ui} = 1\right] = \frac{W^{(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}}$$

is one of the family of models for measurement first described by Rasch in 1960 and that have since come to bear his name.

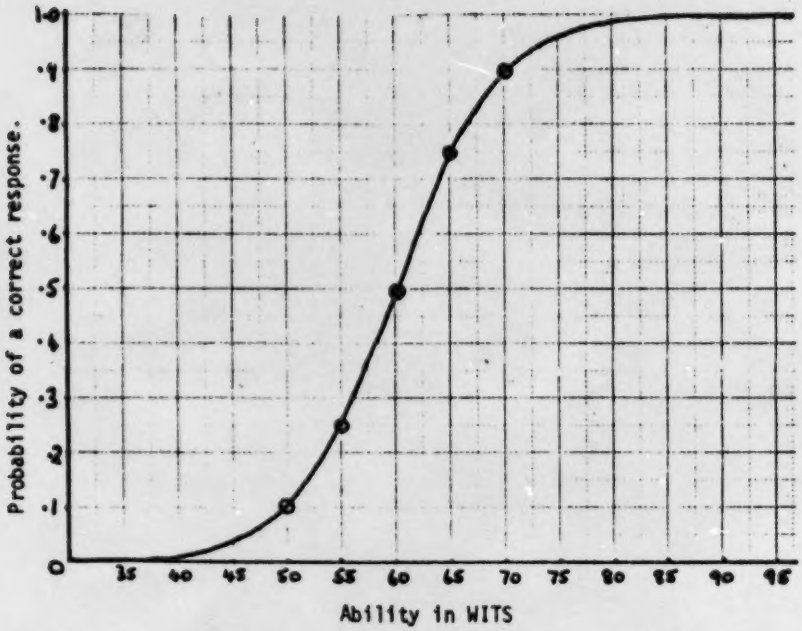
The model proposes that the probability of a particular individual correctly answering a particular test item depends on just two parameters, one measuring the ability of the person and the other the difficulty of the item. Obviously, the ability must be related to the dimensions of difficulty that the item sets. If the item is one on arithmetic, then it is arithmetic ability that is parameterized by α_u, not some generalized ability or ability in some other field of study. Obviously, too, the model is only an approximation to the real situation, and data derived from actual tests do not fit it exactly. It has been found, however, to be quite accurate enough to be useful for a wide range of educational measures, particularly those of achievement.

In applying the model to the analysis of item bank data, the basic problem can be resolved into two separate phases. First, the matrix of accumulated observations (b_{ij}) for the whole bank, or for a part of it, is analysed to yield a set of difficulty parameters (δ_i) for the items. The method used is that of 'maximum likelihood' and the details of the algebraic solution are given in Section 2 of this appendix (Estimation of the parameters). The second phase is the calculation of ability estimates for individuals who answer sub-sets of items drawn from the bank. The method here is essentially the same, though the input information this time consists of the item difficulties as already estimated together with the score (i. e. number of items answered correctly) of the individual concerned. As the Rasch scaled ability in WITS for a particular test depends only on this score and not on which particular items were correctly solved, the ability estimation procedure need be carried out only once for the complete set of possible scores, at the time that a test of items drawn from the bank is constructed.

The particular form of Rasch model chosen (i. e. with W = 1.2457) leads to an especially convenient scale. If we plot an item characteristic curve (i. e. a curve relating the probability of success on the item to the ability of a person attempting it), we obtain the following (this curve is actually for an item with difficulty 60 wits) (see next page for diagram).

A person who has ability 60 wits (i. e. the same as the item's difficulty) has a 50 per cent chance of passing the item.

A person with ability 65 wits (i. e. 5 more than the item parameter) has a 75 per cent chance of success; a person with only 55 wits has 25 per cent chance.



A person with ability 70 wits (i. e. 10 more than the item difficulty) has a 90 per cent chance of success, while somebody of ability 50 wits will succeed on average only 10 per cent of the time. These results, which stem from the fact that W was chosen so that $W^{10} = 9$, offer an easy way to remember relationship between the probability of success, and the discrepancy between a person's ability and an item's difficulty.

2. Estimation of the parameters

For the model to be useful we need a quick, accurate and efficient way of obtaining estimates of difficulty and ability for items and people - given just the data matrix of observations. Several solutions to the problem have been developed though none has succeeded completely to be quick, accurate and efficient. The method described below has the particular advantage that it can be applied to the complex data patterns that arise when not all the students in the sample attempt all the items in the pool. (Such circumstances occur very generally in the operation of an item bank.)

The method, based on the 'relative difficulty' property of items described earlier, is referred to in the literature as the 'pairwise' method, as it examines the data matrix, two items at a time.

The method proceeds in stages. First we use the data to separate our information about each set of parameters - difficulties and abilities. Next we calculate estimates of the difficulties of the test items on the basis of the summarized data and also estimate the standard errors due to sampling fluctuation. Third, we examine a general method for checking the sample-free behaviour of the test items. Fourth, we calculate ability estimates for individuals giving particular sets of responses to the test.

We consider first the results of person u (with ability a_u) attempting two test items i and j (with difficulties δ_i and δ_j). According to the model the probability of success on item i is:

$$\text{Prob}\left(a_{ui} = 1\right) = \frac{W^{(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}} = \frac{W^{\alpha_u}}{W^{\delta_i} + W^{\alpha_u}}$$

and conversely, the probability of failure is:

$$\text{Prob}\left(a_{ui} = 0\right) = \frac{1}{1 + W^{(\alpha_u - \delta_i)}} = \frac{W^{\delta_i}}{W^{\delta_i} + W^{\alpha_u}}$$

Probabilities for the outcome of the encounter with item j are similarly defined, and since the responses to the two items are assumed to be stochastically independent, we can outline four separate cases:

Case (i) - both items correct

$$\text{Prob.}\left(a_{ui} = 1, a_{uj} = 1\right) = \frac{W^{\alpha_u}}{W^{\delta_i} + W^{\alpha_u}} \times \frac{W^{\alpha_u}}{W^{\delta_j} + W^{\alpha_u}}$$

Case (ii) - both items incorrect

$$\text{Prob.}\left(a_{ui} = 0, a_{uj} = 0\right) = \frac{W^{\delta_i}}{W^{\delta_i} + W^{\alpha_u}} \times \frac{W^{\delta_j}}{W^{\delta_j} + W^{\alpha_u}}$$

Case (iii) - Item i correct; item j incorrect

$$\text{Prob.}\left(a_{ui} = 1, a_{uj} = 0\right) = \frac{W^{\alpha_u}}{W^{\delta_i} + W^{\alpha_u}} \times \frac{W^{\delta_j}}{W^{\delta_j} + W^{\alpha_u}}$$

Case (iv) - Item i incorrect; item j correct

$$\text{Prob.}\left(a_{ui} = 0, a_{uj} = 1\right) = \frac{W^{\delta_i}}{W^{\delta_i} + W^{\alpha_u}} \times \frac{W^{\alpha_u}}{W^{\delta_j} + W^{\alpha_u}}$$

The first two cases hold little interest. A moment's reflection reveals that the information they provide about the ability of person u is distinctly limited, and they provide no information at all about the relative difficulties of items i and j.

Cases (iii) and (iv) are somewhat different. If attention is restricted to these two, for both of which $a_{ui} + a_{uj} = 1$, then we can write:

$$\text{Prob.}\left(a_{ui} + a_{uj} = 1\right) = \frac{W^{(\alpha_u + \delta_j)}}{(W^{\delta_i} + W^{\alpha_u}) + (W^{\delta_j} + W^{\alpha_u})} = \frac{W^{(\alpha_u + \delta_i)}}{(W^{\delta_i} + W^{\alpha_u})(W^{\delta_j} + W^{\alpha_u})}$$

$$= \frac{W^{\alpha_u} \cdot (W^{\delta_i} + W^{\delta_j})}{(W^{\delta_i} + W^{\alpha_u})(W^{\delta_j} + W^{\alpha_u})}$$

If therefore we know that person u scored exactly one on the item pair (i, j), then we can write conditional probabilities:

$$\text{Prob.}\left(a_{ui} = 1 \,\middle|\, a_{ui} + a_{uj} = 1\right) = \frac{\dfrac{W^{(\alpha_u + \delta_j)}}{(W^{\alpha_u} + W^{\delta_i})(W^{\alpha_u} + W^{\delta_j})}}{\dfrac{W^{\alpha_u} \cdot (W^{\delta_i} + W^{\delta_j})}{(W^{\alpha_u} + W^{\delta_i})(W^{\alpha_u} + W^{\delta_j})}}$$

$$= \frac{W^{\delta_j}}{W^{\delta_i} + W^{\delta_j}}$$

and similarly:

$$\text{Prob.}\left(a_{uj} = 1 \,\middle|\, a_{ui} + a_{uj} = 1\right) = \frac{W^{\delta_i}}{W^{\delta_i} + W^{\delta_j}}$$

The ability parameter a_u has been eliminated from these two expressions. If we know that an individual scores just one on any item pair, the probability that it was one rather than the other that was answered correctly depends solely on the relative difficulty of the two items.

The probability of getting a correct response to item i, given that of the two responses to items i and j, one is right and one is wrong, can be estimated by observing the results of a large number of people who attempt these two items. If we define b_{ij} to be the number of people who respond correctly to i and incorrectly to j, then we can write:

$$\frac{b_{ij}}{b_{ij} + b_{ji}} \text{ as an estimate of Prob. } \left(a_{ui} = 1 \;\middle|\; a_{ui} + a_{uj} = 1 \right) \text{ or, as above,}$$

$$\frac{b_{ij}}{b_{ij} + b_{ji}} \text{ estimates } \frac{W^{\delta_j}}{W^{\delta_i} + W^{\delta_j}} \text{ and hence } \frac{b_{ij}}{b_{ji}} \text{ estimates } \frac{W^{\delta_j}}{W^{\delta_i}}$$

$$\text{or } (\delta_i - \delta_j) \text{ is estimated by } \frac{\log b_{ji} - \log b_{ij}}{\log W}$$

3. Difficulty wits

To calibrate a set of items from a matrix of observations, a complete B matrix is constructed with elements b_{ij} as defined above. If K items are being considered (i. e. B is of order KxK), then B provides essentially K(K-1)/2 estimation equations for the K unknown parameter δ. Two practical difficulties stand in the way of these equations being used directly. The first is that the equations will almost certainly contain inconsistencies. The second is that particular elements of B may well be zero, so that methods which rely on taking the logs of these entries are not easy to handle computationally.

A satisfactory method of unravelling the information stored in the B matrix is that of maximum likelihood. Suppose that in matrix B, n_{ij} individuals score exactly one on the item pair (i, j), so that we can write:

$$n_{ij} = b_{ij} + b_{ji}$$

Now for any individual in this group, the probability that he gets item i right and item j wrong is:

$$\frac{W^{\delta_j}}{W^{\delta_i} + W^{\delta_j}}$$

and conversely, the probability that i is wrong and j right is:

$$\frac{W^{\delta_i}}{W^{\delta_i} + W^{\delta_j}}$$

Using the binomial theorem, we see that the probability of the n_{ij} individuals dividing into exactly b_{ij} and b_{ji} is:

$$\frac{n_{ij}\,!}{b_{ij}\,!\; b_{ji}\,!} \cdot \frac{W^{b_{ij}\,\delta_j + b_{ji}\,\delta_i}}{(W^{\delta_i} + W^{\delta_j})^{n_{ij}}}$$

The likelihood then of the entire B matrix, given a matrix N of n_{ij} is:

$$\prod_{i,\ j,\ i \neq j} \frac{n_{ij}\ !}{b_{ij}\ !\ b_{ji}\ !} \cdot \frac{W^{(b_{ij}\ \delta_j\ +\ b_{ji}\ \delta_i)}}{(W^{\delta_i} + W^{\delta_j})^{n_{ij}}}$$

The problem now is to find a set of δ_i's which will maximize the value of this function. This maximum and the maximum of its logarithm occur at the same point, so for simplicity we may write the log. likelihood:

$$L = C - \log W. \sum_{i,\ j,\ i \neq j} (b_{ij}\ \delta_j + b_{ji}\ \delta_i) - \sum_{i,\ j,\ i \neq j} (b_{ij} + b_{ji}) \log (W^{\delta_i} + W^{\delta_j})$$

where C is a function of the b's but not of the δ's.

For maximum likelihood $\frac{\partial L}{\partial \delta_i} = 0$ for each i

$$\text{i.e. } 0 = \frac{\partial L}{\partial \delta_i} = \sum^{i} b_{ji} \cdot \log W - \sum^{j} \frac{\log W.\ W^{\delta_i}(b_{ij} + b_{ji})}{(W^{\delta_i} + W^{\delta_j})} \qquad \text{(all i)}$$

$$\text{or } \sum^{i} b_{ij} = \sum^{j} \frac{(b_{ij} + b_{ji}).\ W^{\delta_i}}{(W^{\delta_i} + W^{\delta_j})} \qquad \text{(all i)}$$

This set of K equations in K unknowns can be solved by various iterative techniques. Two that have been found to work well in practice are:

$$\delta_i^{(n+1)} = \frac{\log \left[\sum^{j} b_{ij}\right] - \log \left[\sum^{j} \frac{(b_{ij} + b_{ji})}{(W^{\delta_i} + W^{\delta_j})}\right]}{\log W}$$

and the Newton-Raphson procedure:

$$\delta_{i\,(n+1)} = \delta_i^{(n)} - \frac{\sum^{j} b_{ij} - \sum^{j} \frac{(b_{ij} + b_{ji})\ W^{\delta_i}}{(W^{\delta_i} + W^{\delta_j})}}{\log W. \sum^{j} \frac{(b_{ij} + b_{ji}) W^{(\delta_i + \delta_j)}}{(W^{\delta_i} + W^{\delta_j})^2}}$$

In general an efficient procedure has proved to be to set the initial $\delta^{(0)}$ values all equal to zero, and then to apply the first iterative set of equations three or four times. This produces a set of good approximations which, when used with the Newton-Raphson equations, leads rapidly to convergence, and a solution for the various δ's.

Standard errors for these estimates can now be obtained by calculating:

$$S_i = \sqrt{\frac{-K}{\frac{\partial^2 L}{\partial \delta_i^2}}}$$ for each i (where K is the number of items)

or more accurately from the variance-covariance matrix V where the inverse : V^{-1} has as elements $-\frac{\partial^2 L}{\partial \delta_i \partial \delta_j}$ evaluated at the maximum likelihood solution.

4. Ability wits

If $(a_{ui} ; i = 1, K)$ are the set of scored responses for person u, whose total test scores is r_u, and whose ability parameter is α_u; then we may write:

$$r_u = \sum^{i} a_{ui}$$

The likelihood of the response a_{ui} according to the model is:

$$\frac{W^{a_{ui}(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}}$$

since a_{ui} takes values 1 and zero according as the response is right or wrong.

From this we can see that the likelihood function for the entire set of responses (a_{ui} for person u is:

$$\prod^{i} \left[\frac{W^{a_{ui}(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}} \right]$$

The logarithm of this function:

$$L = \log W. \left[\sum^{i} a_{ui} \alpha_u - \sum^{i} a_{ui} \delta_i \right] - \sum^{i} \log \left[1 + W^{(\alpha_u - \delta_i)} \right]$$

$$= \log W. \left[\alpha_u r_u - \sum^{i} a_{ui} \delta_i \right] - \sum^{i} \log \left[1 + W^{(\alpha_u - \delta_i)} \right]$$

for the ML solution, $\frac{dL}{d\alpha_u} = 0$. It should be noted that in this case the δ's are regarded as already known, so that α_u is the only parameter to be estimated:

$$0 = \frac{dL}{d\alpha_u} = \log W . r_u - \sum^{i} \frac{\log W . W^{(\alpha_u - \delta_i)}}{1+W^{(\alpha_u - \delta_i)}}$$

This equation does not contain the item responses (α_{ui}). It demonstrates a result already obtained by other writers, that the ability estimate depends not upon the particular pattern of item responses obtained, but only upon the 'total score' they yield. Thus the 'total score' r_u is a sufficient statistic for ability, and the conventional practice of using total scores as measures has a logical foundation.

The equation $r_u = \sum^{i} \frac{W^{(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}}$ can be solved for α_u.

The score on the test takes values 0, 1, 2 ... K. Each of the K terms on the right hand side of the equation lies between 0 and 1 for real values of α_u and δ_i. For this reason there are no solutions for r = 0 and r = K. For these values the likelihood function has no maximum, and this could have been anticipated. If an individual responds correctly to every item (i. e. $r_u = K$), then we have no information on which could be based any upper bound for an ability estimate. Similarly, if every item is answered incorrectly, there are no data to suggest just how low the level of ability is.

Note that once a set of items has been calibrated (i. e. the δ's have been estimated), it is possible to estimate an ability parameter for each possible score on the test, regardless of whether or not any individual actually obtains such a score. If a test is constructed by selecting items from an already calibrated pool, then ability parameters for all possible scores on the new test can be calculated even before the test is used.

The standard errors of the ability parameters, corresponding as they do to the standard error of measurement, are usually of more interest than the standard errors of the item difficulties. Furthermore, they are typically considerably larger, since the ability parameter estimates are based upon only K observations (usually between 10 and 100) whereas item calibration is typically based upon the results of at least several hundred individuals.

In general if we assume that the δ's are established with some precision, the standard errors of the α's can be developed from the log. likelihood function.

$$L = \alpha_u r_u \log W - \sum^{i} a_{ui} \delta_i . \log W - \sum^{i} \log \left[1 + W^{(\alpha_u - \delta_i} \right]$$

$$\frac{dL}{d\alpha_u} = \log W.\, r_u - \log W. \sum^{i} \frac{W^{(\alpha_u - \delta_i)}}{1 + W^{(\alpha_u - \delta_i)}}$$

$$\frac{d^2L}{d\alpha_u^2} = -\log W. \sum^{i} \frac{(1 + W^{(\alpha_u - \delta_i)}).\log W.\, W^{(\alpha_u - \delta_i)} - \log W.\, W^{2(\alpha_u - \delta_i)}}{\left[1 + W^{(\alpha_u - \delta_i)}\right]^2}$$

$$= -\left[\log W\right]^2 . \sum^{i} \frac{W^{(\alpha_u - \delta_i)}}{\left[1 + W^{(\alpha_u - \delta_i)}\right]^2}$$

Then the standard error of measurement for an individual who receives an ability estimate α_u for his responses to items with difficulties δ_i is:

$$\frac{1}{\frac{-d^2L}{d\alpha_u^2}}$$

The second differential reaches a maximum value of $\frac{-K(\log W)^2}{4}$ when all the δ's are equal to α_u. In practice if the δ's are all fairly close to α_u (i. e. if the items are all closely matched to the person's ability), then the second differential remains close to

$$\frac{-K\,(\log W)^2}{4}$$

and so the standard error is given approximately by:

$$\frac{2}{\log W.\ \sqrt{K}} \text{ wits}$$

or $\frac{9.10}{\sqrt{K}}$ wits since W = 1.2457

in general, however, we can write

$$S_{\alpha_u} \geq 9/\sqrt{K} \text{ wits}$$

whatever the distribution of the δ's.

APPENDIX 13

DOCUMENTATION FOR AN ITEM BANK

Three separate types of documentation will need to be carefully considered. They are outlined below.

1. Item cards (see Chapter 10 for the format)

A. Purpose of using the items cards : A central store of the texts of items together with summaries of their classification, statistical properties and use.

B. How the cards are arranged and used : The cards will be stored by subject matter and grade level in special filing cabinets. A record will also be kept on each card of when and for what purpose the item was used.

C. Who has access to the cards : Only authorized persons as outlined in (A) above.

2. Users' Manual

A. Purpose of the Manual : to provide necessary information about the bank to test constructors, curriculum developers, research teams, etc.

B. Contents : a. purpose of the item bank;
b. procedures for using the item bank;
c. general details of the classification; scheme for objectives and topics;
d. procedures for constructing tests;
e. summary of sample free scaling theory;
f. notes on interpretation of computer output.

C. Supplement to Manual for each specific subject:
a. analysis of the curriculum;
b. detailed classification scheme for topics;
c. blueprint of items in the bank.

D. Circulation will be strictly limited to those defined in (A) above.

3. Information booklet

A. Purpose of the booklet : to provide teachers, administrators and others who will deal with tests and scores with an understanding of the bank and the units of measurement.

B. Contents : a. the idea of an item bank;
b. guide to interpreting scores when expressed in wits

C. Circulation - not restricted

As described in Chapter 10, the item cards for the storage of the item texts and basic summaries of relevant data have already been introduced for the Grade 11 mathematics test items.

Substantial progress has been made in the drafting of the users' manual, again with the emphasis on Grade 11 mathematics and further sections can be developed from the material presented in Chapter 9 and Appendix 12 of this report.

As yet no attempt has been made to write the information booklet, but this will clearly be needed before any results of item bank based examinations are released to the public.

If the item banking project is to continue, then the development of a complete set of documentation must be given high priorty.

APPENDIX 14

SPECIFICATION OF COMPUTER PROGRAMMES FOR ITEM BANK OPERATION

Programme	Function	Notes
PAIR	1. Calibration of item difficulty 2. Test for sample-free behaviour of test items	Available in IBM/UNIVAC version, to handle up to 90 items in any one test. Could be easily expanded (for IBM 370) to accommodate larger tests (up to say 140 items). The test of sample-free behaviour is accomplished by splitting the sample of students into high and low ability groups, and then comparing item performance on each separately. Graphical printout of item-fit data available for IBM version but not yet incorporated. Due to its demands on storage space, it will not be practicable to adapt program PAIR for use on the WANG system.
TIAP	Performs traditional type of item analysis for test data	Available for IBM/UNIVAC (up to 80 items in a test) and for the WANG system (max. 60 items in a test). IBM version could be easily expanded. Intended for the analysis of multiple-choice data, where each item has one and only one correct response. The test may be treated as a whole or divided into up to 9 sub-tests. Output includes frequency and discrimination indices for every distractor on each item, a matrix (phi-coefficients) of inter-item correlations, a matrix of sub-score correlations, and reliability estimates.
ABLIST	Calibration of test scores in ability wits	IBM/UNIVAC and WANG versions of this programme are both available. The programme is for calibrating scores on new tests composed of items drawn from the bank. Input is the set of calibrated difficulties for the items concerned. Output is a set of ability estimates for the various possible raw scores together with the associated standard errors of measurement.
UPDATE	1. To add new data from tests composed of existing bank items to the accumulated store 2. To add data on new items to the bank 3. To recalibrate the bank 4. To print a directory of bank contents	Written for the IBM 370 system, but not yet tested. Currently structured for banks or sub-banks up to 500 items, but this could be easily extended. For functions 1 and 2 input may be either raw data or a summarized data matrix such as that output by PAIR.